SCALE DEVELOPMENT AND SCORE VALIDATION

EDITED BY: N. Clayton Silver, Laura Badenes-Ribera and Elisa Pedroli

PUBLISHED IN: Frontiers in Psychology and Frontiers in Education

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-782-9
DOI 10.3389/978-2-88963-782-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

SCALE DEVELOPMENT AND SCORE VALIDATION

Topic Editors:
N. Clayton Silver, University of Nevada, Las Vegas, United States
Laura Badenes-Ribera, University of Valencia, Spain
Elisa Pedroli, Italian Auxological Institute (IRCCS), Italy

Citation: Silver, N. C., Badenes-Ribera, L., Pedroli, E., eds. (2020). Scale Development and Score Validation. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88963-782-9

Table of Contents

EDITORIAL
published: 22 April 2020
doi: 10.3389/fpsyg.2020.00799

Editorial: Scale Development and Score Validation

Laura Badenes-Ribera[1]*, *N. Clayton Silver*[2] *and Elisa Pedroli*[3]

[1] Department of Behavioral Sciences Methodology, University of Valencia, Valencia, Spain, [2] Department of Psychology, University of Nevada, Las Vegas, NV, United States, [3] Centro Neuropsicologia, Istituto Auxologico Italiano (IRCCS), Milan, Italy

Keywords: psychological testing, psychometrics, quantitative measurement, questionnaire, scale, reliability, validation

Editorial on the Research Topic

Scale Development and Score Validation

Scale development and validation of scores is not a job to be taken on lightly. Development is a rigorous process which is based on item generation and content validation using expert feedback and pre-testing. In fact, it may take numerous iterations for the scale to be economically feasible and yet convey the appropriate construct.

After the scale has been qualitatively developed, it goes through a rigorous quantitative examination to evaluate its score reliability and validation. This validation may include construct, concurrent, predictive, concurrent, and discriminant. For example, there are numerous techniques for evaluating construct validity such as using exploratory factor analysis (EFA) followed by confirmatory factor analysis (CFA) or using a structural equation model (SEM). Of course, determining the number of factors in an EFA can be quite a problem. Many researchers use the classic Scree test or Kaiser's eigenvalue-greater-than-1.0 technique. However, some studies suggest that these may not be the best techniques (e.g., Lloret-Segura et al., 2014). Other procedures have been developed that allegedly have better psychometric properties, such as Velicer's MAP, parallel analysis, Ruscio and Roche's CD technique, and Achim's NEST method.

Another problem with validation is that the participants are often a single sample (usually college students), which can limit the generalizability of the findings even though cross-validation could still be used. However, we are beginning to witness questionnaires or scales translated into a variety of languages so that factor structures and factor scores become comparable. This cross-cultural work may aid in assessing measurement invariance.

This Research Topic welcomed all types of empirical articles focused on the analysis of the psychometric properties of the measurement instruments in any psychological or social science area. A total of 107 authors contributed 22 articles to the Topic. These articles can be organized intro four issues: (1) Scale development with solid psychometric score validation techniques; (2) Cultural adaptation of developed scales (3) Validation of scores on developed scales, and (4) Invariance measurement of developed scales.

SCALE DEVELOPMENT WITH SOLID PSYCHOMETRIC SCORE VALIDATION TECHNIQUES

Gorostiaga et al. developed and examined the psychometric properties of the Entrepreneurial Orientation Scale (EOS) in a sample of undergraduate students. The EOS showed good

Edited and reviewed by:
*Dominique Makowski,
Nanyang Technological
University, Singapore*

****Correspondence:***
*Laura Badenes-Ribera
laura.badenes@uv.es*

Specialty section:
*This article was submitted to
Quantitative Psychology and
Measurement,
a section of the journal
Frontiers in Psychology*

Received: *28 February 2020*
Accepted: *31 March 2020*
Published: *22 April 2020*

Citation:
*Badenes-Ribera L, Silver NC and
Pedroli E (2020) Editorial: Scale
Development and Score Validation.
Front. Psychol. 11:799.
doi: 10.3389/fpsyg.2020.00799*

psychometric properties and its dimensions demonstrated concurrent relationships with self-efficacy and personal initiative. The EOS may be used to measure entrepreneurial orientation in the educational context and to evaluate interventions designed to promote an entrepreneurial spirit in schools, colleges, and universities.

Shek et al. developed and examined the psychometric properties of the Short form Service Leadership Behavior Scale (SLB-SF-38). This scale was based on the Service Leadership Model proposed by Po Chung. Both EFA and CFA were involved in the validation study. The SLB-SF-38 showed excellent internal consistency, concurrent validity, and factorial validity based on multigroup invariance analyses. The SLB-SF-38 may be used to measure service leadership behavior in the education, research, and personnel training contexts.

Wang D. et al. developed and examined the psychometric properties of a new instrument for depression under the framework of Cognitive Diagnosis Models (CDMs), referred to as CDMs-D. The CDMs-D, which showed good reliability and validity, measures all ten symptom criteria for depression defined in ICD-10 (World Health Organization, 2010) and covers five domains of depression defined by Gibbons et al. (2012). It can also provide both overall information on the severity of depressive disorders and assessment information on specific symptoms defined in the ICD-10, which could be useful for diagnostic and interventional purposes.

Wang J. et al. constructed and validated an instrument to measure psychological security in the area of urban residents' lives known as the Urban Residents Psychological Security Scale (URPS), which showed good reliability and validity using EFA and CFA. This scale can be used as an effective measurement tool for urban residents' psychological security and could be useful for better understanding of residents' demands and monitoring the implementation effects of policies.

Wingenbach et al. created and validated the Verbal Emotion Vignettes as stimulus set to elicit emotions (anger, disgust, fear, sadness, happiness, gratitude, guilt, and neutral) in Portuguese, English, and German. Hierarchical cluster analyses showed that the vignettes mapped clearly on their target emotion categories in all three languages. The final stimulus sets each include 4 vignettes per emotion category plus 1 additional vignette per emotion category, which can be used for task familiarization procedures in research. The high agreement rates on the experienced emotion in combination with the medium-to-large intensity ratings in all three languages suggest that the stimulus sets are suitable for application in emotion research (e.g., emotion recognition or emotion elicitation).

Zhang et al. developed and examined the psychometric properties of the Short-Form Inventory of Callous-Unemotional Traits (ICU, Essau et al., 2006, Chinese version of the ICU: Wang et al., 2017), which was designed to evaluate multiple facets of Callous-Unemotional traits in youths. The short form of the ICU with two factors and 11 items had the best model fit ICU in a Chinese male juvenile offender sample. Both the total and two factor scores showed acceptable internal consistence and convergent validity. The ICU-11 is a promising tool for assessing CU traits in the Chinese male detained juvenile sample.

CULTURAL ADAPTATION OF DEVELOPED SCALES

Rizzo et al. developed the Italian version of the Existential Quest Scale (EQ) and examined factorial structure, internal consistency, discriminant validity, and measurement invariance across gender and age groups. CFA showed that the original one-factor structure was replicated, except for one-item that was removed from the subsequent analyses. Both the internal consistency of the eight-item scale as assessed by Cronbach's and discriminant validity were in line with those of the original study. Furthermore, they found evidence of full measurement invariance across gender and partial measurement invariance across age. Overall, the Italian version of the EQ is a promising tool for assessing flexibility on existential issues.

Ronzón-Tirado et al. adapted the Modified Version of the Conflict Tactics Scale [M-CTS (Neidig, 1986); Spanish adaptation: (Muñoz-Rivas et al., 2007)], in Mexican adolescents using an analysis of the linguistic and cultural variables, followed by a CFA, and the evaluation of Construct and Known Groups Validities. They culturally modified six items and verified the four-factorial structure of the questionnaire. The cultural adaptation of the M-CTS offered adequate reliability and validity scores and expanded the possibilities of comparing the prevalence of the problem between nations with a reliable instrument based on the same theoretical and methodological perspectives.

Yan et al. developed and examined the psychometric properties of the Chinese version of the Brief version of the Situational Test of Emotional Understanding (STEU-B) and the Brief version of the Situational Test of Emotional Understanding (STEM-B) (Allen et al., 2014, 2015) using the Item Response Theory method and criterion validity. The Chinese versions of the STEU-B and STEM-B scales showed psychometrically adequate measurements. These scales might be useful to capture employees' emotional understanding and emotional regulation as an alternative to ability tests of Emotional Intelligence.

VALIDATION OF SCORES ON DEVELOPED SCALES

Angel et al. examined the psychometric properties of the Enriched Life Scale (ELS, Team Red White Blue, 2017) developed to systematically capture and quantify the experiences of military veterans transitioning to civilian life. They used CFA to validate the factorial structure of the ELS in veterans and provided evidence of internal consistence, discriminant, and convergent validity. The ELS could be used in conjunction with diagnostic instruments that capture strain-related transition challenges (to include mental health disorders) to capture post-military service well-being.

Fung et al. assessed the dimensionality and psychometric properties of the Brief Self-Control Scale (BSCS, Chinese version Unger et al., 2016) in a sample of undergraduates using EFA and CFA. A shortened version of the 11-item BSCS with a four-factor structure had better psychometric properties and a good model fit in the CFA. This scale provides a comprehensive and handy measure for broader research in the context of mainland China or the Chinese diaspora.

Tindall and Curtis evaluated the factorial structure of the Need Satisfaction and Frustration Scale (NSFS; Longo et al., 2016) and its predictive validity in a sample of undergraduate students and individuals from the wider community using an SEM. They provided support to Longo et al. (2016, 2018), who stated that need frustration and need satisfaction are distinct constructs, and also gave further insight into the relationship between basic Need Frustration and common types of psychological health problems.

Willmer et al. examined psychometric properties of the 9-item Utrecht work engagement scale (UWES-9, Schaufeli et al., 2006) in a multi-occupational female sample using EFA and CFA. The EFA seemed to mainly favor a one-factor solution, which was shown to explain over 70% of the variance, but none of three different (one-, two-, and three-factor) models showed an overall good fit in CFA. Further research is needed to disentangle the possible effects of gender, nationality, and occupation on work engagement.

Xiao et al. examined the association between student-level information and communication technology (ICT) impact factors (the availability, use and attitudes toward ICT) and reading proficiency among early adolescents using a multiple linear regression model. They found that the students' ICT-related attitudinal factors concerning their interest in ICT and perceived autonomy in using it, rather than its availability and use, were closely associated with high reading proficiency.

ANALYZING THE MEASUREMENT INVARIANCE OF DEVELOPED SCALES

Dagnall et al. evaluated the scale's factorial structure of the Belief in Science Scale (BISS), which assesses the degree to which science is valued as a source of superior knowledge using parallel analysis, EFA, CFA, and invariance testing across gender. They found support to invariance of form, factor structure, and item intercepts for a one-factor model. The scale showed good internal consistency and one-factor solution, signifying that this was consistent with the single-factor model advocated by Farias et al. (2013).

Frey-Clark et al. determined that scores on the Statistical Anxiety Scale (SAS, O'Bryant, 2017) manifest in the same way for students in online and traditional statistics courses using a measurement invariance test.

Martí-Vilar et al. examined the invariance of the Prosocial Behavior Scale (PS, Caprara et al., 2005) across gender and country and psychometric properties in three Hispanic countries (Argentina, Spain, and Peru) using SEM methodology. They also evaluated reliability and internal consistency at both score and item level.

Meng et al. evaluated the factorial structure of the 10-item Connor-Davidson Resilience Scale (CD-RISC-10) in the Chinese elders using CFA and the measurement invariance across gender using multigroup CFA. They found that a single-factor model fitted CD-RISC-10 data well, both for the total sample and for each gender group. Factorial invariance across genders was also supported.

Vagos et al. evaluated the factorial structure of the Morningness-Eveningness-Stability-Scale (MESSi) using CFA and measurement invariance across gender and age using multigroup CFA. They found a three-factor structure for the MESSi and full measurement invariance of the three-factor model for gender and age.

Zhao et al. determined the factor structure of the 15-item Geriatric Depression Scale (GDS-15) in a sample of Chinese elders using CFA and the measurement invariance across gender using multigroup CFA. They found that a three-factor model best fits the structure of the GDS-15, and that measurement invariance across gender was supported, fully assuming different degrees of invariance.

On the other hand, recent developments in statistics have provided new analytical tools for assessing the validity of the scales. French et al. conducted a simulation study to examine the performance of the Generalized Mantel-Haenszel (GMH) procedure and a Multilevel GMH (MGMH) procedure for the detection of uniform differential item functioning (DIF) in the presence of multilevel data with polytomous items. They found differences in DIF detection when the analytic strategy matches the data structure. The GMH had an in?ated Type I error rate across conditions and thus an artificially high power rate, and the MGMH had good power rates while maintaining control of the Type I error rate. Finally, Hayduk et al. detailed the relevant procedural steps to conduct a fusion validity and illustrated the procedure using the Leadership scale from the Alberta Context Tool (ACT) with care aides working in Canadian long-term care homes.

This Research Topic includes different examples of scale development and validation protocols, each one with rigor and scientific peculiarity. We had analyzed four different aspects of this wide field of knowledge: scale development with solid psychometric score validation techniques, cultural adaptation of developed scales, validation of scores on developed scales, and invariance measurement of developed scales. It's important to show how variegate these processes could be with the aim of promote the use of different scientific-based techniques.

AUTHOR CONTRIBUTIONS

LB-R, EP, and NS all helped in writing the editorial.

ACKNOWLEDGMENTS

The editors greatly appreciate the contributions received from the authors on this Research Topic.

REFERENCES

Allen, V., Rahman, N., Weissman, A., MacCann, C., Lewis, C., and Roberts, R. D. (2015). The situational test of emotional management-brief (STEMB): development and validation using item response theory and latent class analysis. *Pers. Individ. Diff.* 81, 195–200. doi: 10.1016/j.paid.2015.01.053

Allen, V. D., Weissman, A., Hellwig, S., MacCann, C., and Roberts, R. D. (2014). Development of the situational test of emotional understanding-brief (STEU-B) using item response theory. *Pers. Individ. Diff.* 65, 3–7. doi: 10.1016/j.paid.2014.01.051

Caprara, G., Steca, P., Zelli, A., and Capanna, C. (2005). A new scale for measuring adults? prosocialness. *Eur. J. Psychol. Assess.* 21, 77–89. doi: 10.1027/1015-5759.21.2.77

Essau, C. A., Sasagawa, S., and Frick, P. J. (2006). Callous-unemotional traits in community sample of adolescents. *Assessment* 13, 454–469. doi: 10.1177/1073191106287354

Farias, M., Newheiser, A. K., Kahane, G., and de Toledo, Z. (2013). Scientific faith: belief in science increases in the face of stress and existential anxiety. *J. Exp. Soc. Psychol.* 49, 1210–1213. doi: 10.1016/j.jesp.2013.05.008

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., et al. (2012). Development of a computerized adaptive test for depression. *Arch. Gen. Psychiatry* 69, 1104–1112. doi: 10.1001/archgenpsychiatry.2012.14

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., and Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada [Exploratory item factor analysis: a practical guide revised and updated]. *Anal. Psicol.* 30, 1151–1169. doi: 10.6018/analesps.30.3.199361

Longo, Y., Alcaraz-Ibáñez, M., and Sicilia, A. (2018). Evidence supporting need satisfaction and frustration as two distinguishable constructs. *Psicothema* 30, 74–81. doi: 10.7334/psicothema2016.367

Longo, Y., Gunz, A., Curtis, G. J., and Farsides, T. (2016). Measuring need satisfaction and frustration in educational and work contexts: the need satisfaction and frustration scale (NSFS). *J. Happiness Stud.* 17, 295–317. doi: 10.1007/s10902-014-9595-3

Muñoz-Rivas, M. J., Andreu, J. M., Graña, J. L., O'Leary, K. D., and González, M. P. (2007). Validation of the modified version of the Conflicts tactics scale (M-CTS) in a Spanish population of youths. *Psicothema* 19, 693–698.

Neidig, P. M. (1986). *The Modified Conflict Tactics Scale*. Beaufort, SC: Behavioral Sciences Associates.

O'Bryant, M. J. (2017). *How attitudes towards statistics courses and the field of statistics predicts statistics anxiety among undergraduate social science majors: a validation of the Statistical Anxiety Scale* (Doctoral dissertation). ProQuest LLC; University of North Texas. Available online at: https://search.proquest.com/docview/2009455494

Schaufeli, W. B., Bakker, A. B., and Salanova, M. (2006). The measurement of work engagement with a short questionnaire a cross-national study. *Educ. Psychol. Meas.* 66,701–716. doi: 10.1177/0013164405282471

Team Red White Blue (2017). *The Enriched Life Scale*. Tampa, FL: Team Red, White & Blue.

Unger, A., Bi, C., Xiao, Y.-Y., and Ybarra, O. (2016). The revising of the Tangney Self-Control Scale for Chinese students. *PsyCh J.,* 5, 101–116. doi: 10.1002/pchj.128

Wang, M.-C., Gao, Y., Deng, J., Lai, H., Deng, Q., and Armour, C. (2017). The factor structure and construct validity of the inventory of callous-unemotional traits in Chinese undergraduate students. *PLoS ONE* 12:e0189003. doi: 10.1371/journal.pone.0189003

World Health Organization (2010). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Psychology

ORIGINAL RESEARCH
published: 17 January 2019
doi: 10.3389/fpsyg.2019.00003

Factorial Structure of the Morningness-Eveningness-Stability-Scale (MESSi) and Sex and Age Invariance

*Paula Vagos[1,2], Pedro F. S. Rodrigues[3], Josefa N. S. Pandeirada[3,4], Ali Kasaeian[5], Corina Weidenauer[5], Carlos F. Silva[3,4] and Christoph Randler[5]**

[1] INPP, Universidade Portucalense, Porto, Portugal, [2] CINEICC, University of Coimbra, Coimbra, Portugal, [3] CINTESIS, Department of Education and Psychology, University of Aveiro, Aveiro, Portugal, [4] William James Research Center, University of Aveiro, Aveiro, Portugal, [5] Department of Biology, Eberhard Karls University of Tübingen, Tübingen, Germany

Assessing morningness-eveningness preferences (chronotype), an individual characteristic that is mirrored in daily mental and physiological fluctuations, is crucial given their overarching influence in a variety of domains. The current work aimed to investigate the best factor structure of an instrument recently presented to asses this characteristic: the Morningness-Eveningness-Stability-Scale improved (MESSi). For the first time, the originally proposed three-factor structure was pitched against a uni- and a two-factor solution. Another novelty was to establish that the best-fitting model would be invariant in relation to sex and age, two variables that influence chronotype. A Confirmatory Factor Analyses on the data obtained from a sample of 2096 German adults (age: 18–76; *M* = 25.5, *SD* = 7.64) revealed that the originally proposed three-factor structure of the MESSi – Morning Affect, Eveningness, and Distinctness – was the only one to achieve acceptable fit indicators. Furthermore, each scale obtained good internal consistency. In order to assess age invariance, following the literature on development and chronotype, our sample was divided into three age groups: 18–21 years, 22–31 years, and 32 years or older. Full measurement invariance of the three-factor model was found for sex and age. Regarding differences between sexes, females did not differ significantly from males in Morning Affect, but scored significantly lower on Eveningness and higher on Distinctness; this last result has been consistent across validation studies of the MESSi. With respect to age differences, the oldest group scored lower on Eveningness and Distinctness in comparison with the other two age-groups; the intermediate group (age: 22–31) scored lower on Morning Affect when compared to both the younger and older age groups. Additionally, both Eveningness and Distinctness were negatively correlated with age. This latter relation has been consistently reported in other validation studies. Our results reinforce the idea that the MESSi assesses three different components of chronotype in a reliable manner and that this instrument can be used to explore sex and age differences.

Keywords: MESSi, three-factor structure, sex invariance, age-group invariance, distinctness, morning affect, eveningness, psychometric assessment

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Arcady A. Putilov,
Humboldt-Universität zu Berlin,
Germany
Vincenzo Natale,
University of Bologna, Italy

****Correspondence:***
Christoph Randler
christoph.randler@uni-tuebingen.de

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *13 November 2018*
Accepted: *03 January 2019*
Published: *17 January 2019*

Citation:
Vagos P, Rodrigues PFS, Pandeirada JNS, Kasaeian A, Weidenauer C, Silva CF and Randler C (2019) Factorial Structure of the Morningness-Eveningness-Stability-Scale (MESSi) and Sex and Age Invariance. Front. Psychol. 10:3. doi: 10.3389/fpsyg.2019.00003

INTRODUCTION

People differ in the time of the day in which the peak of mental and physiological functions occurs (chronotype) and can be classified in one of three types: morning-, evening-, or intermediate-types. Specifically, whereas in morning-types the peak of alertness arises in early hours, in evening-types it occurs in the afternoon/evening; the peak of intermediate-types is reached in the middle of the day (Schmidt et al., 2007; Adan et al., 2012). Concerning body temperature, the nadir occurs at 03:50 h in morning-types and at 06:01 h for evening-types (Baehr et al., 2000). This individual difference is relevant in a variety of domains. For example, it has been related to affective conditions (e.g., Randler et al., 2012; Oginska and Oginska-Bruchal, 2014), to health-related behaviors and problems (e.g., Fabbian et al., 2016; Suh et al., 2017), and to satisfaction with life (e.g., Randler, 2008; Jankowski, 2012). Chronotype also relates in different ways to various characteristics of personality (e.g., Lipnevich et al., 2017; Randler et al., 2017b). These examples justify the need to seriously consider this variable in research in an accurate manner (for a review, see also Adan et al., 2012).

Although chronotype can be assessed by different biological and objective methods (e.g., melatonin, body temperature and actimetry measurements), self-report questionnaires continue to be widely used (for a review, see Di Milia et al., 2013). Some examples are the Morningness-Eveningness Questionnaire (full form-MEQ, Horne and Östberg, 1976; reduced form-rMEQ, Adan and Almirall, 1991) or the Composite Scale of Morningness (CSM; Smith et al., 1989). More recently, Randler et al. (2016a) proposed another instrument to assess circadian preferences – the Morningness-Eveningness-Stability-Scale improved (MESSi) – that includes three subscales: Morning Affect, Eveningness, and Distinctness. Alike other instruments, the Morning Affect and Eveningness subscales indicate more morningness and eveningness preference, respectively. The Distinctness subscale measures the subjective amplitude or the range of fluctuations that occur during the day in the mental and physiological state of the individual. Whereas some individuals present a relatively stable state throughout the day (i.e., they do not feel strong differences in their state during the day), others experience larger variations (i.e., they perceive to be doing particularly well at some point in the day and worse in others); the first are considered to have a low amplitude and the later a high amplitude (Oginska, 2011; for related concepts, see also Folkard et al., 1979; Di Milia, 2005; Oginska et al., 2017).

The MESSi provides several improvements in relation to previous questionnaires (Di Milia et al., 2013; Randler et al., 2016a). For example, it includes a similar number of items formulated to assess morning and eveningness preferences, thus avoiding the morning-biased measurement characteristic of other instruments. It also clearly identifies the assessment of multiple dimensions. Even though previous instruments have been proposed to assess multi-dimensions of chronotype (e.g., Putilov, 1993; Roberts, 1998), and factor analysis exist on other morningness-eveningness scales (Neubauer, 1992; Brown, 1993; Caci et al., 2009), the MESSi suggests a novel three-factor structure. The wording of the items of the MESSi is also more updated and the questions are simpler to respond and interpret. Finally, the inclusion of the Distinctness, a dimension with growing recognized relevance in the assessment of circadian rhythm (Di Milia, 2005; Oginska, 2011; Dosseville et al., 2013), makes it a more complete instrument, which of course goes on charge of the length. Nevertheless, in comparison to other popular alternatives, the MESSi (composed of 15 items) adds a new dimension and still provides a shorter solution than the MEQ (composed of 19 items); as compared to the CSM (which contains 13 items) it only adds two items.

The MESSi has been submitted to several validation studies, namely in Germany, Spain, Iran, Portugal, and Slovenia (Randler et al., 2016a; Díaz-Morales and Randler, 2017; Diaz-Morales et al., 2017; Rahafar et al., 2017; Rodrigues et al., 2018; Tomažič and Randler, 2019). In short, all studies have replicated the three-factor internal structure (i.e., Morning Affect, Eveningness, and Distinctness) via exploratory (Randler et al., 2016a) or confirmatory factor analyses (Díaz-Morales and Randler, 2017; Diaz-Morales et al., 2017; Rahafar et al., 2017; Rodrigues et al., 2018). However, the factor structure has not been challenged by comparing a one-, two- or three-factor structure. These validation studies showed at least satisfactory internal consistency values (Cronbach' alphas varying between 0.73 and 0.87 for Morning Affect, 0.80 and 0.84 for Eveningness, and 0.69 and 0.77 for Distinctness). Rahafar et al. (2017) further found the MESSi to be invariant at the configuration level only across the three countries involved in their study (Germany, Spain. and Iran); in other words, the three-factor model fitted acceptably for each country but the loadings and intercepts of items (particularly for the Eveningness measure) seem to differ across countries. Furthermore, Rodrigues et al. (2018) found evidence for strong invariance of the MESSi across men and women in a Portuguese sample of higher education students. Finally, though not explicitly testing for measurement invariance, Diaz-Morales et al. (2017) showed the three-factor model to acceptably fit different age groups (i.e., 17–30 years old and 31–65 years old). Therefore, testing factorial invariance is an important novel goal of this study.

Concurrent validity of the MESSi has also been confirmed against other typical questionnaires. Specifically, Morning Affect correlated positively and Eveningness correlated negatively with the CSM (Randler et al., 2016a) and with the rMEQ (Díaz-Morales and Randler, 2017; Faßl et al., 2018). Regarding Distinctness, the correlation between its scores and the CSM and the rMEQ was negative but lower than with the other two subscales (Randler et al., 2016a; Díaz-Morales and Randler, 2017). Moreover, in the study by Faßl et al. (2018), no correlations were found between Distinctness and the other subscales. Overall, these results suggest that Distinctness acts separately from Morning Affect and Eveningness. These authors also reported some preliminary evidence for the MESSi chronotype assessment using measures of actigraphy and of the sleep-wake rhythm.

The literature on circadian preferences has also explored how these change throughout the development and if there are differences between sexes. Studies that have assessed chronotype using the MESSi, have revealed inconsistent sex differences on Morning Affect and Eveningness (e.g., Díaz-Morales and Randler, 2017; Diaz-Morales et al., 2017; Rahafar et al., 2017). This inconsistency mimics that obtained when other instruments are used to asses chronotype and may be a result of (low) sample size and high variation in age (Randler, 2007; Adan et al., 2012). Regarding the subscale of Distinctness, the results have been very regular across all of the just mentioned studies, with females reporting higher Distinctness than males (e.g., Rahafar et al., 2017; Rodrigues et al., 2018).

The evaluation of chronotype in different age groups has revealed that children tend to be morning-oriented and then become more evening-oriented during adolescence (e.g., Roenneberg et al., 2004; Randler et al., 2017a). Morningness usually increases again, particularly after the age of 20/21 years, and tends to stabilize until individuals reach around the age of 30 (Roenneberg et al., 2004; Adan et al., 2012; Randler et al., 2016b). Some of the studies that have used the MESSi have reported positive relations between Morning Affect and age and negative relations between Distinctness and age (e.g., Díaz-Morales and Randler, 2017; Rodrigues et al., 2018). Regarding the relation between Eveningness and age, the results have been more irregular, with some reporting negative relations (e.g., Díaz-Morales and Randler, 2017; and some countries from the Rahafar et al., 2017 study) and others non-significant relations (Rodrigues et al., 2018).

Given the existing literature, the main aim of the current work was to test competing models for the factorial structure of the MESSi and the invariance across age classes and sex of the best fitting model. In other words, the current work aimed to test the originally proposed three-factor structure of the MESSi (Morning Affect, Eveningness, and Distinctness) against uni- and two-factor model solutions. The first comparison helps to establish the multidimensionality purpose that underlined the development of this instrument (Randler et al., 2016a). The second evaluation aims to explore the idea that morningness-eveningness corresponds to a single dimension (Di Milia and Randler, 2013; Diaz-Morales et al., 2017) that in turn differs from the dimension of Distinctness. Furthermore, we aimed to establish that the best-fitting model would be invariant in relation to sex and age. This is an important statistical procedure in psychometric research to assure comparability across the groups being considered (Schmitt and Ali, 2015). With the exception of the study by Rodrigues et al. (2018), no other validation study of the MESSi has directly investigated the invariance of its factorial structure concerning sex and no other study has looked at the invariance for age groups. Finally, we also explored the differences between sexes and among age groups in the scores of each subscale of the MESSi (Morning Affect, Eveningness, and Distinctness).

MATERIALS AND METHODS

Sample

Participants were 2096 adults aged between 18 and 76 years (M = 25.5, SD = 7.64); two participants did not provide information on their age (0.1%). The majority of participants was female (n = 1458, 69.6% females; n = 619, 29.5% males); nineteen participants (0.9%) did not provide information on their sex. Men were significantly older than women (M = 26.51, SD = 8.65 and M = 25.03, SD = 7.06, respectively, $t(980.79) = 3.76$, $p < 0.001$). For data analysis purposes (see below), participants were divided into three age groups: 21 years old or younger (n = 693, 33%), 22–31 years old (n = 1127, 54%), and 32 years old or older (n = 276, 13%). Such division took into account some of the ages at which stronger changes in chronotype are expected to occur (c.f. Introduction) while also ensuring a reasonable number of participants per age group. Men and women were not evenly distributed by these age groups, $\chi^2(2) = 9.04$, $p = 0.01$, with men being overrepresented in the two younger groups and women being more prevalent in the older group, as compared to what was statistically expected.

Instrument

The MESSi is a self-report instrument that includes 15 items from three other questionnaires. The original items are from the Composite Scale of Morningness (Smith et al., 1989), the Caen Chronotype Questionnaire (CCTQ, Dosseville et al., 2013) and the Circadian Energy Scale (CIRENS; Ottoni et al., 2011). The total of the items is divided in three subscales, each one composed of five items: Morning Affect, Eveningness, and Distinctness. The items related to the Morning Affect subscale measure morningness preferences (early schedules), whereas the items of the Eveningness subscale assess evening preferences (late schedules). The remaining five items constitute the Distinctness subscale, that is, the amplitude dimension of this instrument. Each item is responded using a 5-points Likert scale and scored with 1–5 points, although some of them are reverse coded. The previous validation studies mentioned in the Introduction have revealed good indexes, such as Cronbach' alpha values for the three subscales ranging between 0.69 to 0.87.

Procedure

Sampling and Data Collection

Data collection was done from 23.10.2017 until 13.11.2017. Students and employees of the Eberhard Karls University of Tübingen were contacted by e-mail and asked to participate in a study about sleep and sexual behavior. In that same e-mail they were informed that it was a short questionnaire study about chronotype and partnership and that it would last about 15 min. They were also told that an anonymized procedure was in place, that their data would be used only for research purposes, and that they could withdraw their participation at any time without any consequences. We also explicitly stated that it was a voluntary and unpaid study. Then, participants were directed to a website from "SoSci Survey" where they had to answer to the questions; the consent of the participants was implied by completing the questionnaire. The questions concerning the MESSi took approximately 5 min to complete. We did not control for double or triple access. Two participants were excluded from the sample due to being under 18 years of age.

Data Analyses

A Confirmatory Factor Analyses (CFA) approach was used to test for competing models that might underlie the internal structure of the MESSi. Three measurement models were tested: (1) a one-factor model including all 15 items; (2) a two-factor model considering a Morning Affect/Eveningness factor with 10 items and a Distinctness factor with 5 items; and (3) a three-factor model referring to a Morning Affect factor, an Eveningness factor, and a Distinctness factor, each with five items. For the two-factor model, the scoring of the items from the Eveningness scale were reversed turning them into items contributing to a Morningness evaluation as if we were dealing with a morning-eveningness continuum (rather than two separate subscales as initially intended). The fit of these models was judged based on the guidelines provided by Hair et al. (2014) for samples larger than 250 participants and instruments using between 12 and 30 items. Therefore, the models were considered to fit the data if showing comparative fit index (CFI) > 0.92 combined with standardized root mean square residual (SRMR) < 0.08 or with root mean square error of approximation (RMSEA) < 0.07. Only one of the tested models acceptably fitted the data (see results section) and so only its measurement invariance by sex and by age-groups was analyzed, based on a forward approach (Dimitrov, 2010). Firstly, configural invariance was established if the model was found to fit well within each group under analyses. Then, metric invariance was investigated, meaning that the model that constraints all loadings to be equal across groups should be as good a fit as the model posing no equality constraints on the groups (i.e., ΔCFI < −0.01; ΔSRMR < 0.03; ΔRMSEA < 0.03). Finally, scalar invariance was also tested, based on finding a non-expressive difference between the loading-constraint model and a model constraining all intercepts to be equal across groups (i.e., ΔCFI < −0.01; ΔSRMR < 0.03; ΔRMSEA < 0.01; Chen, 2007).

Following the establishment of measurement invariance, a latent mean comparison approach was taken for between and among group comparisons (i.e., sex and age-groups, respectively). These analyses were further complemented with effect sizes, descriptive data and a two between-factor ANOVA to control for the uneven distribution of men and women by age-groups. These last analyses, as well as the calculations of the Cronbach's alpha as a measure for internal consistency, were carried out using the IBM SPSS Statistics 21. In turn, CFA, measurement invariance, latent mean comparisons, between factor correlation analyses and correlation analyses between subscales and age were ran using Mplus v7.4 (Muthén and Muthén, 2012).

RESULTS

Preliminary analysis showed the data on the 15 items of the MESSi for the 2096 participants were not multivariate normal (Mardia's multivariate skewness statistic = 6.59, $p < 0.001$; Mardia's multivariate kurtosis statistic = 281.42, $p < 0.001$; Korkmaz et al., 2014). Hence, and because there were no missing values, the Robust Maximum Likelihood estimator was used for confirmatory factor analyses and for measurement analyses. Also, non-parametric tests were used for the correlation analyses.

Evidence Based on the Internal Structure of the MESSi

The three factor measurement model originally proposed for the MESSi (Randler et al., 2016a) was the only one to achieve acceptable fit indicators based on the combination between CFI and SRMR values; the one-factor and the two-factor solutions did not abide by the fit guidelines for any of the indices under consideration (c.f. **Table 1**). All three measures also achieved mostly good internal consistency values: α = 0.87 for Morning Affect, α = 0.85 for Eveningness, and α = 0.75 for Distinctness. Loading values were always significant and varied between 0.65 (CSM 4) and 0.84 (CCQ 4) for Morning Affect, between 0.44 (CCQ 11) and 0.91 (CCQ 2) for Eveningness, and between 0.46 (CCQ 6) and 0.72 (CCQ 15) for Distinctness (c.f. **Supplementary Material**). The Morning Affect scale correlated significantly ($p < 0.001$) and negatively with the Eveningness ($r = −0.59$) and the Distinctness ($r = −0.38$) scales; Eveningness and Distinctness were also positive and significantly correlated although at a borderline significance level and with a low correlation value ($r = 0.06$, $p = 0.041$).

Full measurement invariance by sex was established for the three-factor model given that it fitted well for female and male participants taken separately (i.e., configural invariance; c.f. **Table 1**)[1], that forcing all item loadings to be equal between groups did not significantly worsened the fit of a non-constraint model (i.e., metric invariance; ΔCFI = 0.000, ΔRMSEA = −0.002 and ΔSRMR = 0.002), and, additionally, that forcing all item intercepts to be equal across groups again did not significantly worsened the fit of the loading constraint model (i.e., scalar invariance; ΔCFI = −0.004, ΔRMSEA = 0.000, and ΔSRMR = 0.003)[2].

Evidence for the three levels of measurement invariance by age-groups was also found, namely configural invariance (c.f. **Table 1**)[3], metric invariance (ΔCFI = 0.000, ΔRMSEA = −0.003,

[1] Loading values for female participants varied between 0.45 (CCQ 6 and CCQ 11) and 0.92 (CCQ 2; c.f. **Supplementary Material**) and internal consistency values were 0.85 for Morning Affect, 0.86 for Eveningness and 0.75 for Distinctness. Loading values for male participants ranged from 0.39 (CCQ 11) to 0.89 (CCQ 2; c.f. **Supplementary Material**) and internal consistency values were 0.86 for Morning Affect, 0.81 for Eveningness and 0.73 for Distinctness.

[2] The same results were attained when randomly selecting a subsample of 50% of the female sample ($n = 702$) to contrast with the complete male sample ($n = 619$). That proportion was chosen so that the male and female groups had a similar size. Further information on the results using this sample may be requested from the corresponding author.

[3] Loading values for participants aged 21 years old or younger varied between 0.42 (CCQ 6) and 0.89 (CCQ 2; c.f. **Supplementary Material**) and internal consistency values were 0.85 for Morning Affect, 0.83 for Eveningness and 0.71 for Distinctness. Loading values for participants aged between 22 and 31 years ranged from 0.44 (CCQ 11) to 0.92 (CCQ 2; c.f. **Supplementary Material**) and internal consistency values were 0.88 for Morning Affect, 0.85 for Eveningness and 0.75 for Distinctness. As for the participants aged 32 years old or older, loading values were placed between 0.41 (CCQ 11) and 0.94 (CCQ 2; c.f. **Supplementary Material**) and

TABLE 1 | Confirmatory factor analyses on the internal structure of the MESSi.

	χ2	df	RMSEA	CI for RMSEA	CFI	SRMR
Confirmatory factor analyses						
1 Factor measurement model	5539.60	90	0.170	0.166; 0.174	0.553	0.134
2 Factor measurement model	4101.11	89	0.147	0.143; 0.151	0.671	0.096
3 Factor measurement model	993.69	87	0.071	0.067; 0.74	0.926	0.053
Between-sex measurement invariance						
Female participants	747.03	87	0.072	0.067; 0.077	0.927	0.054
Male participants	338.45	87	0.068	0.061; 0.076	0.919	0.053
Between-age-groups measurement invariance						
21 years old or younger	383.11	87	0.070	0.063; 0.077	0.917	0.055
Between 22 and 31 years old	622.02	87	0.074	0.068; 0.079	0.920	0.058
31 years old or older	191.77	87	0.066	0.053; 0.079	0.946	0.060

df, degrees of freedom; RMSEA, root mean square error of approximation; CI, confidence interval; CFI, comparative fit index; SRMR, standardized root mean square residual. All chi-square values were significant at p < 0.001.

and ΔSRMR = 0.003), and scalar invariance (ΔCFI = −0.002, ΔRMSEA = −0.002, and ΔSRMR = 0.001)[4].

Between-Groups Comparisons

Latent mean comparisons indicate that women, compared to men, scored significantly lower on the Eveningness (latent mean = −0.029, $p < 0.001$) and significantly higher on the Distinctness scale (latent mean = 0.563, $p < 0.001$); scores on the Morning Affect scale did not differ significantly between sexes. The direction of these results reflect those found for the same measures and groups when taking the sum of the responses of the set of items composing each measure (c.f. **Table 2**, also for the descriptive measures found using the complete sample).

Concerning age, correlation analyses revealed that age correlated positively with Morning Affect (r = 0.08, p = 0.003) and negatively with Eveningness (r = −0.08, $p < 0.001$) and Distinctness (r = −0.125, $p < 0.001$). Furthermore, latent mean comparisons showed that the oldest group had the lowest scores on the Eveningness and Distinctness scales, compared to both the younger group (latent mean = −0.182, p = 0.012 and latent mean = −0.145, p = 0.038, respectively) and the group of participants aged 22–31 years old (latent mean = −0.269, $p < 0.001$ and latent mean = −0.281, $p < 0.001$, respectively). In turn, participants aged between 22 and 31 years had significantly lower scores on the Morning Affect when compared to the younger group (latent mean = −0.152, p = 0.002) and to the older group (latent mean = 0.217, p = 0.002). The direction of these results, again, is in line with that found for the same measures and groups when taking the sum of the responses of the set of items composing each scale (c.f. **Table 2**).

Because men and women were not evenly distributed by age-groups, we conducted an ANOVA including both age-groups and sex as between groups factors. Their interaction effect was non-significant for the Morning Affect [F(2,2076) = 2.308,

TABLE 2 | Descriptive measures for the subscales of the MESSi and effect sizes for between-groups comparisons.

	Morning affect				Eveningness				Distinctness			
	M (SD)	25th	50th	75th	M (SD)	25th	50th	75th	M (SD)	25th	50th	75th
Total sample	15.54 (4.27)	13	16	19	15.83 (4.11)	13	16	19	17.06 (3.34)	15	17	19
Sex												
Female	15.44 (4.29)	13	16	19	15.48 (4.2)	12	15	19	17.55 (3.19)	16	18	20
Male	15.75 (4.19)	13	16	19	16.68 (3.76)	14	17	19	15.92 (3.41)	14	16	18
Cohen's *d*	0.07				0.30				0.49			
Age-groups												
21 years old or younger	15.85 (3.94)	13	16	19	16.03 (3.94)	13	16	19	17.18 (3.28)	13	18	19
Between 22 and 31 years old	15.19 (4.35)	12	15	19	15.89 (4.15)	13	16	19	17.19 (3.28)	15	18	19
32 years old or older	16.18 (4.68)	13	16	20	15.11 (4.29)	12	15	18	16.21 (4.64)	14	16	19
Partial eta-squared	0.008				0.005				0.010			

internal consistency values were 0.89 for Morning Affect, 0.87 for Eveningness and 0.81 for Distinctness.

[4] The same results were attained when randomly selecting a subsample of 33% of the participants aged 21 years old or younger (n = 232) and a subsample of 25% of the participants aged 22–31 years old (n = 305) to contrast with the complete sample of participants aged 32 years or older (n = 276). Those proportions were chosen to make group sizes as similar as possible. Further information on the results using this sample may be requested from the corresponding author.

$p = 0.10$], for the Eveningness, and for the Distinctness (both *F*s < 1). These results suggest that sex- and age-based differences on the MESSi seem to be independent of each other.

DISCUSSION

The MESSi provides new way of assessing circadian preferences while introducing several improvements as compared to other existing instruments. Here, we tested the originally proposed three-factor structure of the MESSi (Morning Affect, Eveningness, and Distinctness), against other possible factorial structures. Also, we assessed the factor invariance across age groups and sex. The current study addressed these novel issues using a large sample of participants. Our results confirmed that the originally proposed three-factor structure of the instrument provides a better fit to the data as compared to the alternatives of a one- and two-factor structure.

Some studies that have tested the concurrent validity of the MESSi against other instruments (e.g., MEQ) have found correlations of about the same size as ours (but of different direction) between both the Morning Affect and Eveningness (Diaz-Morales et al., 2017; Rodrigues et al., 2018); such results could suggest that morningness-eveningness is a unidimensional construct and not separate as proposed in the MESSi (Diaz-Morales et al., 2017). However, our results suggest that each of the three different factors contribute separately to the assessment of chronotype. Empirically, studies have further started to show that each of these dimensions relate in a differential and significant manner with health-related measures as well as with some personality characteristics (Diaz-Morales et al., 2017) which helps to establish the relevance of each of the three factors. Furthermore, each scale obtained good internal consistency (range 0.75–0.87) scores.

The correlations found among the subscales are in line with those reported in other studies. The correlations between Morning Affect and both Eveningness and Distinctness were negative and significant with a larger relation between the first two, as expected (Díaz-Morales and Randler, 2017; Rodrigues et al., 2018). The correlation between Distinctness and Eveningness was also significant but with a low positive correlation coefficient; a similar result was reported by Rodrigues et al. (2018) but others have revealed non-significant correlations (Diaz-Morales et al., 2017).

Establishing that the best-found model would be invariant for the variables of sex and age was also an important and novel goal of this work. Full measurement invariance of the three-factor model was obtained for these variables indicating that the MESSi can accurately reflect sex and age differences related to the constructs. Such results reassure researchers that the MESSi accurately grasps the constructs within sex- and age- diversified samples and is an appropriate instrument to compare the results between sexes and across age groups.

We also explored the differences between sexes and among age groups in the scores of each subscale of the MESSi. Even though our sample was composed of unequal groups per sex or age, the same results were obtained when using balanced-sized groups (see footnotes 2 and 4). The pattern of differences between sexes has been quite inconsistent across studies, particularly with respect to the dimensions of Morning Affect and Eveningness, but we were able to find some communality with our data. Specifically, our females scored lower than males on Eveningness and the difference was not significant for Morning Affect (Diaz-Morales et al., 2017, undergraduate sample; Rodrigues et al., 2018). On the other hand, the finding that females score higher on Distinctness than males has been more consistently reported (e.g., Rahafar et al., 2017).

Regarding age, our correlation results revealed that as participants get older, they tend to score lower on Eveningness and Distinctness and higher on Morning Affect. This last result is in agreement with the idea that after the end of adolescence, people tend to become more morning oriented (Roenneberg et al., 2004), a relation that has also been corroborated in other studies using the MESSi (Díaz-Morales and Randler, 2017; Rahafar et al., 2017; Rodrigues et al., 2018). On the other hand, the negative correlation between Eveningness and age has been replicated in some studies (e.g., Diaz-Morales et al., 2017) but not in others (Rodrigues et al., 2018; the correlation was negative but non-significant). The negative correlation between age and Distinctness obtained in our sample has also been found in most validation studies of the MESSi in which this relation was analyzed (e.g., Rahafar et al., 2017; Rodrigues et al., 2018). Note that the disparate results regarding the correlations between age and Morning Affect and Eveningness are in favor of the idea that the latter two are indeed different constructs. Finally, we found no significant interaction between age and sex, a result that differs from that reported by Diaz-Morales et al. (2017). As for the differences among the age groups, considering the scarceness of studies that have addressed them before, we refrain from discussing these data at this time.

The diversity of results regarding the relation between age and the three subscales of this instrument could be due to a number of factors such as the different age ranges that have been tested across studies and the differential sample sizes. Furthermore, there is a number of factors that seem to affect chronotype such as individual and environmental variables (e.g., age, sex and photoperiod at birth, longitude and altitude; Adan et al., 2012); consequently, one could expect variability across countries as these differ in many of these aspects. It is noteworthy, though, that some results have indeed been consistent such as finding that females score consistently higher on Distinctness than males and the negative correlation between age and Eveningness and Distinctness. Future studies should explore the factors likely underlying these consistencies and also those that might justify the discrepancies.

In sum, this study confirms that the best fitting model for our data include the three factors described in the original presentation of the MESSi: Morning Affect, Eveningness and Distinctness. We further demonstrated that such structure is invariant for the variables of sex and age which ensures researchers that all of the instrument can be reliably used to assess chronotype in males and females as well as in various age groups. We also provide additional information regarding the relation between these two variables and chronotype in our sample with contributes to a more global understanding of this variable across countries.

AUTHOR CONTRIBUTIONS

CR, AK, and CW designed the study and collected the data. PV, PFSR, JNSP, and CFS made the analyses and drafted the manuscript. All authors contributed to the writing and discussion and approved the manuscript.

FUNDING

This research was supported by Gips-Schüle-Stiftung. We acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of University of Tübingen.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00003/full#supplementary-material

REFERENCES

Adan, A., and Almirall, H. (1991). Horne & östberg morningness-eveningness questionnaire: a reduced scale. *Pers. Individ. Dif.* 12, 241–253. doi: 10.1016/0191-8869(91)90110-W

Adan, A., Archer, S. N., Hidalgo, M. P., Milia, L. D., Natale, V., and Randler, C. (2012). Circadian typology: a comprehensive review. *Chronobiol. Int.* 29, 1153–1175. doi: 10.3109/07420528.2012.719971

Baehr, E. K., Revelle, W., and Eastman, C. I. (2000). Individual differences in the phase and amplitude of the human circadian temperature rhythm: with an emphasis on morningness-eveningness. *J. Sleep Res.* 9, 117–127. doi: 10.1046/j.1365-2869.2000.00196.x

Brown, F. M. (1993). Psychometric equivalence of an improved Basic Language Morningness (BALM) scale using industrial population within comparisons. *Ergonomics* 36, 191–197. doi: 10.1080/00140139308967872

Caci, H., Deschaux, O., Adan, A., and Natale, V. (2009). Comparing three morningness scales: age and gender effects, structure and cut-off criteria. *Sleep Med.* 10, 240–245. doi: 10.1016/j.sleep.2008.01.007

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

Di Milia, L. (2005). A psychometric evaluation and validation of the preferences scale. *Chronobiol. Int.* 22, 679–693. doi: 10.1080/07420520500180454

Di Milia, L., Adan, A., Natale, V., and Randler, C. (2013). Reviewing the psychometric properties of contemporary circadian typology measures. *Chronobiol. Int.* 30, 1261–1271. doi: 10.3109/07420528.2013.817415

Di Milia, L., and Randler, C. (2013). The stability of the morning affect scale across age and gender. *Pers. Individ. Diff.* 54, 298–301. doi: 10.1016/j.paid.2012.08.031

Díaz-Morales, J. F., and Randler, C. (2017). Spanish adaptation of the Morningness-Eveningness-Stability-Scale (MESSi). *Span. J. Psychol.* 20:e23. doi: 10.1017/sjp.2017.21

Diaz-Morales, J. F., Randler, C., Arrona-Palacios, A., and Adan, A. (2017). Validation of the MESSi among adult workers and young students: general health and personality correlates. *Chronobiol. Int.* 34, 1288–1299. doi: 10.1080/07420528.2017.1361437

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Meas. Eval. Couns. Dev.* 43, 121–149. doi: 10.1177/0748175610373459

Dosseville, F., Laborde, S., and Lericollais, R. (2013). Validation of a chronotype questionnaire including an amplitude dimension. *Chronobiol. Int.* 30, 639–648. doi: 10.3109/07420528.2012.763042

Fabbian, F., Zucchi, B., De Giorgi, A., Tiseo, R., Boari, B., Salmi, R., et al. (2016). Chronotype, gender and general health. *Chronobiol. Int.* 33, 863–882. doi: 10.1080/07420528.2016.1176927

Faßl, C., Quante, M., Mariani, S., and Randler, C. (2018). Preliminary findings for the validity of the Morningness-Eveningness-Stability Scale improved (MESSi): correlations with activity levels and personality. *Chronobiol. Int.* doi: 10.1080/07420528.2018.1519570 [Epub ahead of print].

Folkard, S., Monk, T. H., and Lobuan, M. C. (1979). Towards a predictive test of adjustment to shift work. *Ergonomics* 22, 79–91. doi: 10.1080/00140137908924591

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2014). *Multivariate Data Analysis*. Upper Saddle River: Prentice Hall.

Horne, J., and Östberg, O. (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int. J. Chronobiol.* 4, 97–110.

Jankowski, K. S. (2012). Morningness/eveningness and satisfaction with life in a Polish sample. *Chronobiol. Int.* 29, 780–785. doi: 10.3109/07420528.2012.685671

Korkmaz, S., Goksuluk, D., and Zararsiz, G. (2014). MVN: an r package for assessing multivariate normality. *R J.* 6, 151–162.

Lipnevich, A. A., Crede, M., Hahn, E., Spinath, F. M., Roberts, R. D., and Preckel, F. (2017). How distinctive are morningness and eveningness from the big five factors of personality? A meta-analytic investigation. *J. Pers. Soc. Psychol.* 112, 491–509. doi: 10.1037/pspp0000099

Muthén, L. K., and Muthén, B. O. (2012). *Mplus User's Guide*. Los Angeles, CA: Muthén.

Neubauer, A. C. (1992). Psychometric properties of two circadian rhythm questionnaires and their relationship with personality. *Pers. Individ. Differ.* 13, 125–132. doi: 10.1016/0191-8869(92)90035-N

Oginska, H. (2011). Can you feel the rhythm? A short questionnaire to describe two dimensions of chronotype. *Pers. Individ. Diff.* 50, 1039–1043. doi: 10.1016/j.paid.2011.01.020

Oginska, H., Mojsa-Kaja, J., and Mairesse, O. (2017). Chronotype description: in search of a solid subjective amplitude scale. *Chronobiol. Int.* 34, 1388–1400. doi: 10.1080/07420528.2017.1372469

Oginska, H., and Oginska-Bruchal, K. (2014). Chronotype and personality factors of predisposition to seasonal affective disorder. *Chronobiol. Int.* 31, 523–531. doi: 10.3109/07420528.2013.874355

Ottoni, G. L., Antoniolli, E., and Lara, D. R. (2011). The circadian energy scale (CIRENS): two simple questions for a reliable chronotype measurement based on energy. *Chronobiol. Int.* 28, 229–237. doi: 10.3109/07420528.2011.553696

Putilov, A. A. (1993). "A questionnaire for self-assessment of individual profile and adaptability of sleepwake cycle," in *Chronobiology and Chronomedicine 1991: Basic Research and Applications*, eds C. Gutenbrunner, G. Hildebrandt, and R. Moog (Frankfurt am Main: Lang), 492–498.

Rahafar, A., Randler, C., Díaz-Morales, J. F., Kasaeian, A., and Heidari, Z. (2017). Cross-cultural validity of morningness-eveningness stability scale improved (MESSi) in Iran, Spain and Germany. *Chronobiol. Int.* 34, 273–279. doi: 10.1080/07420528.2016.1267187

Randler, C. (2007). Gender differences in morningness–eveningness assessed by self-report questionnaires: a meta-analysis. *Pers. Individ. Diff.* 43, 1667–1675. doi: 10.1016/j.paid.2007.05.004

Randler, C. (2008). Morningness-eveningness and satisfaction with life. *Soc. Indicat. Res.* 86, 297–302. doi: 10.1007/s11205-007-9139-x

Randler, C., Díaz-Morales, J. F., Rahafar, A., and Vollmer, C. (2016a). Morningness-eveningness and amplitude - development and validation of an improved composite scale to measure circadian preference and stability (MESSi). *Chronobiol. Int.* 33, 832–848. doi: 10.3109/07420528.2016.1171233

Randler, C., Freyth-Weber, K., Rahafar, A., Florez Jurado, A., and Kriegs, J. O. (2016b). Morningness-eveningness in a large sample of German adolescents and adults. *Heliyon* 2:e00200. doi: 10.1016/j.heliyon.2016.e00200

Randler, C., Faßl, C., and Kalb, N. (2017a). From lark to owl: developmental changes in morningness-eveningness from new-borns to early adulthood. *Sci. Rep.* 7:45874. doi: 10.1038/srep45874

Randler, C., Schredl, M., and Göritz, A. S. (2017b). Chronotype, sleep behavior, and the big five personality factors. *SAGE Open* 7:2158244017728321. doi: 10.1177/2158244017728321

Randler, C., Stadler, L., Vollmer, C., and Diaz-Morales, J. F. (2012). Relationship between depressive symptoms and sleep duration/chronotype in women. *J. Individ. Diff.* 33, 186–191. doi: 10.1027/1614-0001/a000089

Roberts, R. D. (1998). *The Lark-Owl (chronotype) Indicator (LOCI)*. Sydney: Entelligent Testing Products.

Rodrigues, P. F. S., Vagos, P., Pandeirada, J. N. S., Marinho, P. I., Randler, C., and Silva, C. F. (2018). Initial psychometric characterization for the portuguese version of the morningness-eveningness-stability-Scale improved (MESSi). *Chronobiol. Int.* 35, 1608–1618. doi: 10.1080/07420528.2018.1495646

Roenneberg, T., Kuehnle, T., Pramstaller, P. P., Ricken, J., Havel, M., Guth, A., et al. (2004). A marker for the end of adolescence. *Curr. Biol.* 14, R1038–R1039. doi: 10.1016/j.cub.2004.11.039

Schmidt, C., Collette, F., Cajochen, C., and Peigneux, P. (2007). A time to think: circadian rhythms in human cognition. *Cogn. Neuropsychol.* 24, 755–789. doi: 10.1080/02643290701754158

Schmitt, N., and Ali, A. A. (2015). "The practical importance of measurement invariance," in *More Statistical and Methodological Myths and Urban Legends*, eds C. E. Lance and R. J. Vandenberg (New York, NY: Routledge), 327–346.

Smith, C. S., Reilly, C., and Midkiff, K. (1989). Evaluation of three circadian rhythm questionnaires with suggestions for an improved measure of morningness. *J. Appl. Psychol.* 74, 728–738. doi: 10.1037/0021-9010.74.5.728

Suh, S., Yang, H.-C., Kim, N., Yu, J. H., Choi, S., Yun, C.-H., et al. (2017). Chronotype differences in health behaviors and health-related quality of life: a population-based study among aged and older adults. *Behav. Sleep Med.* 15, 361–376. doi: 10.1080/15402002.2016.1141768

Tomažič, I., and Randler, C. (2019). Slovenian adaptation of the Morningness-Eveningness-Stability Scales improved (MESSi). *Biol. Rhythm Res.* (in press) doi: 10.1080/09291016.2018.1535539.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 March 2019
doi: 10.3389/fpsyg.2019.00619

Cultural Adaptation of the Modified Version of the Conflicts Tactics Scale (M-CTS) in Mexican Adolescents

Rosa Carolina Ronzón-Tirado, Marina Julia Muñoz-Rivas, María Dolores Zamarrón Cassinello and Natalia Redondo Rodríguez*

Department of Biological and Health Psychology, Universidad Autónoma de Madrid, Madrid, Spain

Several scales are used in Dating Violence studies assuming cross-cultural invariance and equivalence of the measures without making the proper validation in the intended populations. This study focuses on the importance of adapting existing dating violence psychological instruments (as the widely recognized Modified Version of the Conflict Tactics Scale, M-CTS) in diverse adolescent populations adjusting to international validation procedures that ensure the cultural fit of the instrument and the measurement invariance of the construct. We sought to adapt the M-CTS in Mexican adolescents (*N* = 1861; 57.5% woman) following the ITC Guidelines for Translating and Adapting Test. We made an analysis of the linguistic and cultural variables, followed by a Confirmatory Factor Analysis, and the evaluation of Construct and Known Groups Validities. We culturally modified six items and verified the four-factorial structure of the questionnaire proposed in previous studies (argumentation, psychological aggression, mild physical aggression, and sever physical aggression). We also found significant correlations in between the scores of the M-CTS and the Aggression Questionnaire (AQ) and the Dominating and Jealous Tactics Scale (DJTS), verifying the Construct Validity of the M-CTS to measure aggressive behaviors. Conclusion: the cultural adaptation of the M-CTS offered adequate reliability and validity scores in Mexican population expanding the possibilities of comparing prevalences of the problem between nations with a reliable instrument based on the same theoretical and methodological perspectives.

Keywords: dating violence, psychological testing, validity, cultural adaptation, Mexican adolescents

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Cesar Merino-Soto,
Universidad de San Martín de Porres, Peru
Amelia Rizzo,
University of Messina, Italy

****Correspondence:***
Rosa Carolina Ronzón-Tirado
rosa.ronzon@estudiante.uam.es

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *27 November 2018*
Accepted: *06 March 2019*
Published: *21 March 2019*

Citation:
Ronzón-Tirado RC, Muñoz-Rivas MJ, Zamarrón Cassinello MD and Redondo Rodríguez N (2019) Cultural Adaptation of the Modified Version of the Conflicts Tactics Scale (M-CTS) in Mexican Adolescents. Front. Psychol. 10:619. doi: 10.3389/fpsyg.2019.00619

INTRODUCTION

Psychometric test are not always adapted properly before they are used within two different cultures (Gjersing et al., 2010; Borsa et al., 2012). Researchers usually change test instructions, response formats, or the number and content of the items without taking into account if the modifications are suitable for the new context or consistent with the original version. Although these are probably well-intention actions based on the strong psychometric properties of the original instruments, they end up compromising the quality of the results (Eremenco et al., 2005; Reichenheim and Moraes, 2007).

Aware of this lack of rigor in the use of measurement tools, organizations such as the American Educational Research Association, the American Psychological Association, the European Federation of Psychologist Association, and the International Test Commission have generated guidelines in the last two decades for the development, administration, validation, and psychometric tests adaptation. Specifically, since 1976, the ITC has focused its efforts on the validation process (Oakland et al., 2009; Muñiz et al., 2015) and has edited a specific journal on the subject since 1998 (Hambleton and Patsula, 1999). It has also published the *ITC Guidelines for Translating and Adapting Tests* in International Test Commission [ITC] (2005), and its version 2.4 (2016), which main object has been to stablish a reliable method to cross-culturally adapt, administrate, and interpret tests.

Despite these important advances in the adaptation field, the most widely used scales still those specifically developed for the English-speaking population (Byrne and Van de Vijver, 2010; Muñiz et al., 2013). Testing the scales' psychometric properties in other cultures or countries is necessary for the progress of research in topics that had been widely recognized as public health concerns (World Health Organization [WHO], 2002) such as teen dating violence.

During the last 5 years, there has been an increase in descriptive dating violence studies in Latino American cultures (Rodríguez, 2014; Celis-Sauce and Rojas-Solís, 2015; Boira et al., 2017; Rey-Anacona et al., 2017; Rojas-Solís et al., 2017). These studies, however, have not focused on using instruments adapted to the intended populations, making comparisons between groups difficult and hindering more concluding results. Specifically in Mexico, a remarkable variability has been found in the prevalence of documented aggressions during dating relationships, ranging from 46 to 86% of cases (Peña-Cárdenas et al., 2013; Carrillo-Flores, 2014; Vega-Valero, 2015; Oliva-Zárate et al., 2018). The available data is not conclusive and differs in terms of the theoretical models and methodologies used, as well as in the selection of the measurement instruments, which are generally created ex professor for each case and which psychometric properties are not usually reported.

In addition, it should be noted that the documented prevalence of teen dating violence in Mexico, as in other countries, has mainly been carried out in a global manner, without analyzing the directionality of the different behavioral expressions of the aggressions (Rubio-Garay et al., 2012). Few studies have discriminated the experiences of victimization/perpetration or have differentiated between verbal aggressions, mild physical aggressions and severe physical aggressions. Therefore, validate internationally recognized measurement instruments of dating violence, is an important contribution to recognize the magnitude of the problem and its characteristics, as well as for the development of prevention programs and intervention of violence in relationships in the Latin American context (Fernández-Fuentes et al., 2011; Fernández-González et al., 2013; Rubio-Garay et al., 2017).

Among the most widely used instruments for measuring teen dating violence in Latino America, the modified version (Cascardi et al., 1999) of the M-CTS (Neidig, 1986), stands out as one of the most appropriate scales to respond to the current demand for cross-cultural and multilingual evaluation of the problem (Ryan, 2013). This, unlike other scales, has shown adequate psychometric properties in previous adaptations in the United States (Straus, 2004), Italy (Nocentini et al., 2011), and Spain (Muñoz-Rivas et al., 2007a).

Although, the M-CTS has already been validated in Spanish-speaking population (Muñoz-Rivas et al., 2007a), there is still a lack of adaptations for Latin American countries. It would be a mistake to assume the permanence of the psychometric guarantees of the Spain validation in the rest of the Spanish-speaking countries. Applying the M-CTS without taking into account cultural variables between nations, could imply that the data obtained do not really reflect the reality of the adolescents, but the discrepancy in the understanding of the teen dating violence mediated by cultural and temporal variables such as religion, lifestyle and values. As well as, discrepancies originated by physical characteristics of the M-CTS like the item format and material of the test (Gjersing et al., 2010; International Test Commission [ITC], 2016).

For example, Latinos are said to hold more traditional attitudes about women, relationships and commitment, and Mexicans may have more rigid expectations about gender roles than North American or European populations. Although this kind of believes are changing and may vary across urban and rural groups, the powerful subjective influence of these believes over dating violence measure most be recognized (Hokoda et al., 2006; Shaffer et al., 2018).

In addition, when performing cross-cultural comparative studies, the variants found may not show the similarities or differences between countries, but the deficiencies of the M-CTS when evaluating each population mediated by the use of the language, such as, family structure of the language or semantic equivalence (Eremenco et al., 2005). Ryan et al. (1999) for example, found a lack of measurement equivalence when they attempted to apply attitudes surveys in a multinational organization where Spanish and Mexican employees worked. To reduce the lack of invariance they needed to make two Spanish versions of the surveys. After the adjustments, the wording of the items of each version clearly differed although the items represented similar content.

The objective of this study was to adapt the M-CTS Spanish version (Muñoz-Rivas et al., 2007a) in Mexican adolescents following internationally accepted guidelines proposed by International Test Commission [ITC] (2016). We hypothesize (a) to confirm the reliability and validity of the adapted M-CTS to measure different types of aggression in Mexican teen dating relationships. (b) that the cultural adaptation of the M-CTS would maintain the four-factor structure proposed in previous validations; (c) that the cultural adaptation of the M-CTS could discriminate different scores based on sex and age of the respondents; and that (d) that the M-CTS would correlate significantly with other scales that measure general aggression such as Aggression Questionnaire (AQ; Buss and Perry, 1992) and psychological violence in adolescents such as the Dominating and Jealous Tactics Scale (DJTS, Kasian and Painter, 1992).

MATERIALS AND METHODS

Participants

The sample comprised 1,861 adolescents from six public schools in Xalapa (Veracruz, México). Inclusion criteria were (a) having had or currently having a dating relationship, (b) being between 12 and 18 years old (c) fluent Spanish reading and understanding (d) not presenting developmental disabilities incompatible with the requirements of the survey administration. 57.5% were women and 42.5% men, with a mean age of 15.5 years (*SD* = 1.39, range = 12–18), 47.6% of them were early adolescents (ages 12–15) and 52.4% late adolescents (ages 16–18). While 38% of the participants reported having a dating relationship with an average duration of 9.25 months (*SD* = 10.4), 62% reported not dating anyone currently but having done before (*M* = 5.82 months, *SD* = 7). The 91% reported having a heterosexual orientation, 7.1% bisexual, and 1.9% homosexual. Data was collected by convenience sampling method during the 2017–2018 school period.

Instruments

Participants completed a questionnaire composed of sociodemographic and dating relationships data, as well as the instruments listed below:

The Modified Conflict Tactics Scale (M-CTS; Neidig, 1986) Spanish adaptation (Muñoz-Rivas et al., 2007a), is made up of 18 bidirectional items with a 5-point response format, ranging from 1 (*never*) to 5 (*very often*), assesses perpetration and victimization of psychological and physical violence. The answer frame of the question refers to the current relationship or last one in the case that the respondent do not have a relationship by the survey moment. It has a four-factor structure (i.e., argumentation; psychological violence; mild physical violence; and severe physical violence); and, in the Spanish adaptation, reliability, measured through Cronbach's alpha coefficient in the subscales of Aggression, ranged from 0.65 to 0.82 for Perpetration and from 0.63 to 0.82 for Victimization (Muñoz-Rivas et al., 2007a). Scores interpretation: all the items have the same direction, each punctuation of the 8 subscales, indicates whether the respondent has been involved in such conduct, such as the frequency of the aggression in the reference period. The individual items can be examined together with the total scores of the subscales by the different implications that they could have, as an example, give a slap in comparison with punching.

The *Dominating and Jealous Tactics Scale* (DJTS; Kasian and Painter, 1992), Spanish validation (Muñoz-Rivas et al., 2019) has been used to analyze the convergent validity of M-CTS in measuring perpetration and victimization of psychological violence in courtship. It is made up of 11 bidirectional items with a 5-point response (from 1 "never" to 5 "very frequently") to measure perpetration and victimization of dominant and jealous tactics. In the Spanish adaptation the reliability of the scale was good for both perpetration and victimization (Cronbach α = 0.76 and α = 0.78, respectively; Muñoz-Rivas et al., 2019). In the present sample, the result of the Exploratory Factor Analysis indicated that the eleven items, for both perpetration and for victimization scales were distributed in two factors (Dominant and Jealous tactics), the total variance explained by the two factors in the perpetration model was 38.1%, and 41.85% for the victimization model. The reliability of the perpetration scale was α = 0.77 and α = 0.82 for victimization scale, whit α-values for the domination and jealous scales between 0.67 and 0.79.

The *Aggression Questionnaire* (AQ; Buss and Perry, 1992), Spanish version (Andreu et al., 2002) is comprised of 29 Likert-type items with five response options (from 1 "totally agree" to 5 "totally disagree") grouped into four factors: physical aggression (α = 0.86), verbal aggression (α = 0.86), anger (α = 0.86), and hostility (α = 0.86). It has been used in order to evaluate the convergent validity of the M-CTS to measure levels of general aggressiveness. In the present sample, the AQ scale obtained an Exploratory Analysis of the AQ Scale indicated, as in the Spanish validation, that the 29 items were distributed in 4 factors (physical aggression, verbal aggression, anger and hostility). The total variance explained by the 4 factors were 38,61%. The reliability of the verbal aggression scale was α = 0.68, α = 0.76 for physical aggression scale, α = 0.72 for anger scale, and α = 0.77 hostility.

Although DJTS and AQ have not been adapted yet to Mexican adolescents, they have been used to test convergent validity of the M-CTS in this study due to: (a) the lack of adapted Mexican scales to measure this constructs (López-Cepero et al., 2015) and, (b) their proven strong psychometric properties in English-speaking and Spanish young adults and adolescents samples (Cascardi et al., 1999; Muñoz-Rivas et al., 2007b, 2009; Chaín-Pinzón et al., 2012; Cascardi and Avery-Leaf, 2015).

Procedure

The methodology proposed in the *ITC Guidelines for Translating and Adapting Test* (International Test Commission [ITC], 2016) was followed to carry out the adaptation. Guidelines and procedural objectives are reflected in **Table 1**.

The questionnaire was administered during school hours with prior informed consent of the participants, their parents, and the school's supervisors and principals. Before the administration, the researchers provided participants information about the aims of the research, procedures, confidentiality protections, and participants' right to withdraw the study. The classrooms were designated as sample units, and the approximate response time of the questionnaire participants was 50 min. The evaluators were trained in the use of the scale by both the authors of the Spanish version and Mexican researchers.

Descriptive statistics and departure from the normality of the variables were made follow by Exploratory Factor Analyses (EFA) using General Least Square (GLS) method of estimation and reliability test for AQ and DJTS scales (both scales have been used to test the convergent validity of the M-CTS). Afterwards, *Mann–Whitney U test* were performed to asses difference between M-CTS scores by sex and age, effect size was measured with *A static*. Then Spearman correlations were made between subscales to test convergent validity of the M-CTS. All of these analyses were made using the statistical package, SSPS v20 (IBM, 2011).

Finally, the Structural Equation Models were tested using the Mplus 7.0 software (Muthén and Muthén, 1998–2015) Due to the distribution of the variables MLM estimator was used. To study model-fit, the following indexes and values were considered (Jöreskog, 2001; Hooper et al., 2008): Root Mean Square Error of Approximation (Good fit = 0 ≥ RMSEA ≤ 0.05; Acceptable fit = 0.05 ≥ RMSEA ≤ 0.08); Standardized Root Mean Square Residual (Good fit = 0 ≥ RMSEA ≤ 0.05; Acceptable fit = 0.05 ≥ RSMR ≤ 0.1) and Comparative Fit Index

TABLE 1 | Summary of the ITC guidelines for translating and adapting test (2016).

Precondition guidelines

PC-1 (1) Obtain the permission from the intellectual holder of the original scale.

PC-2 (2) Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use.

PC-3 (3) Minimize the influence of any irrelevant cultural and linguistic differences (e.g., religion).

Test development guidelines

TD-1 (4) Ensure that the translation and adaptation process consider linguistic, psychological, and cultural differences in the intended populations (ask experts on the subject).

TD-2 (5) Use appropriate translation designs and procedures to maximize the suitability of the test adaptation. Focus on functional rather than on a literal equivalence.

TD-3 (6) Provide evidence that the test instructions and item content have similar meaning for the intended populations.

TD-4 (7) Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for the intended populations.

TD-5 (8) Collect pilot data on the adapted test to enable item analysis, reliability assessment, and small-scale validity studies. Make any necessary changes.

Confirmation guidelines

C-1 (9) Select sample with characteristics and sufficient size for the intended use and relevance for the empirical analyses.

C-2 (10) Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence.

C-3 (11) Provide evidence supporting the norms, reliability, and validity of the adapted version.

C-4 (12) Use an appropriate equating design and data analysis procedures when linking score scales from different language versions.

Administration guidelines

A-1 (13) Minimize any culture- and language-related problems that are caused by administration procedures and response modes.

A-2 (14) Specify testing conditions that should be followed closely in all interest populations.

Score scales and interpretation guidelines

SSI-1 (15) Interpret any group score differences with reference to all relevant available information.

SSI-2 (16) Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported.

Documentation guidelines

Doc-1 (17) Provide technical documentation of any changes.

Doc-2 (18) Provide documentation for test users that will support good practice in the use of the adapted test in the context of the new population.

(Acceptable Fit = CFI $\geq$ 0.9). Reliability of the M-CTS Subscales was measured using *Cronbach's Alpha* and *Omega* coefficients.

RESULTS

The results obtained for each phase indicated in the ITC Guidelines are described in this section (International Test Commission [ITC], 2016; **Table 1**).

Precondition Guidelines

The license to use the scale was obtained from the authors of the Spanish version of the M-CTS (Muñoz-Rivas et al., 2007a), and researchers obtained the approval of the Research Ethics Committee of the Autonomous University of Madrid to carry out the study (CEI-85-1576). Subsequently, two dating violence experts (i.e., Spanish and Mexican postdoctoral researchers with more than 10 years of experience on the topic and several published studies about dating violence) qualitatively analyzed the instrument to verify the equivalence of the construct and to minimize the influence of cultural variables (e.g., lifestyles and value systems) in both populations. The evaluation was positive, and no modifications were necessary.

Test Development Guidelines

Two independent postdoctoral Mexican researchers, experts in dating violence and skilled in psychometrics, made adaptations to the content of the scale. They focused on grammar, terminology, and the colloquial use of words to ensure that the adaptation process considered the cultural, psychological, and linguistic differences of Mexican adolescents (Borsa et al., 2012). They agreed on the modification of items 6, 8, and 14, (in perpetration and victimization scales). In item 6, "estabais" was replaced by "estaban"; in item 8, "picar" and "picarte" were replaced by "molestar" and "molestarte"; and in item 14, "abofeteado" by "dar una cachetada." Once the scale was modified, the authors of the Spanish version verified that the proposed modifications did not alter the construct.

To empirically support the modifications, a pilot test of the scale was conducted using a sample of 118 adolescents randomly selected from two educational centers in Xalapa. The sample was made up of 50.8% women and 42.2% men with ages between 12 and 17 years (M = 14.81 years; SD = 1.42). The reliability of the scale was analyzed using the Cronbach's Alpha coefficient and Confidence Intervals 95%, in all cases the coefficient provided statistically acceptable scores similar to those obtained in the Spanish version (i.e., α = 0.46 CI [0.26–0.61] and 0.44 CI [0.24–0.60] for argumentation; 0.68 CI [0.58–0.77] and 0.59 CI [0.45–0.69] for verbal aggression; α = 0.81 CI [0.76–0.86] and 0.75 CI [0.68–0.82] for mild physical aggression; and, 0.76 CI [0.68–0.83] and 0.56 CI [0.40–0.68] for severe physical aggression, perpetration and victimization subscales).

In addition, the convergent validity of the test was analyzed using the AQ and DJTS scales. Positive and significant Spearman correlations were found for: (a) The M-CTS psychological violence subscales and DJTS dominant tactics subscales ($r_s = 0.44$, $p < 0.001$, for perpetration; and $r_s = 0.45$, $p < 0.001$, for victimization); (b) The M-CTS Psychological Violence subscales and the DJTS Jealous Tactics subscales ($r_s = 0.49$, $p < 0.001$, for perpetration; and $r_s = 0.48$, $p < 0.001$ for victimization); (c) The MCTS Psychological Violence subscales and the AQ Verbal Aggression subscale ($r_s = 0.20$, $p < 0.001$).

Positive significant Spearman correlations were also found between the subscales of the (a) M-CTS Mild Physical Violence perpetration subscale and the subscale of physical aggression of the AQ ($r_s = 0.17$, $p < 0.001$). There was no significant correlation in-between M-CTS Severe Physical Violence subscale and the AQ Physical Aggression subscale ($r_s = 0.04$, $p = 0.054$), this last result is explained by the items content of both subscales, since the level f aggressiveness is much higher u the items used in the M-CTS.

Confirmation Guidelines

Once the pilot had concluded, the M-CTS was administered to a large sample of 1,861 adolescents from Xalapa. Results follow.

Reliability

The reliability of perpetration and victimization M-CTS subscales was estimated through the Cronbach's Alpha coefficient and the Confidence Intervals 95% (CI 95%) for each case. The CI 95% was estimated to assess the precision of the α measures and determine between what values the α coefficient could oscillate in the population (Domínguez-Lara and Merino-Soto, 2015). The analysis revealed Cronbach's Alpha scores between α = 0.43 for Argumentation on the victimization scale and α = 0.78 for Mild Physical Violence victimization. The coefficients values of Argumentation and Sever Physical aggression subscales were under 0.5 but still acceptable taking into account the scare number of items of each subscale (Crutzen and Ygram, 2017). Additionally, Omega coefficients were also calculated because it has been shown (Ventura-León and Caycho-Rodríguez, 2017) that unlike the coefficient of alpha, Omega provides more precise reliability measures as it works with factorial loads (**Table 2**).

Furthermore, given the importance of this instrument for professional and epidemiological practice, reliability between relevant groups have been calculated. Analysis in early adolescents subgroup reveled acceptable Cronbach's Alpha scores between α = 0.78 [CI 0.32–0.48] for mild physical victimization and α = 0.58 [CI 0.50–0.65] for severe physical victimization, and values of 0.40 [CI 0.32–0.48] and 0.46 [CI 0.35–0.54] for perpetration and victimization argumentation subscale. Analysis in late adolescents reveled acceptable Cronbach's Alpha scores between α = 0.65 [CI 0.63–0.68] for verbal aggression perpetration and α = 0.79 [CI 0.77–0.80] for mild physical victimization, and values of 0.46 [CI 0.41–0.51] and 0.42 [CI 0.37–0.47] for perpetration and victimization argumentation subscale. The results for argumentation subscales still acceptable considering the scare number of the items in each one.

Confirmatory Factor Analysis

Due to the distributions of the variables, the confirmatory factor analysis was conducted using the MLM maximum likelihood parameter with standard errors and a mean-adjusted chi-square test statistic that are robust to non-normality. Compared to de ML estimation, a robust MLM approach is less dependent of the assumption of multivariated normal distribution and have the advantage of computing robust versions of CFI and RMESEA. Thus, the use of MLM estimator was the most appropriate approach for the analysis (Byrne, 2012). The structural equation models were configured according to the four factor structure (for both perpetration and victimization scales) that previous studies had supported in North American (Caulfield and Riggs, 1992; Pan et al., 1994; Straus, 2004) and Spanish samples (Muñoz-Rivas et al., 2007a). Additionally two factor structure proposed by Cascardi et al. (1999) was tested, it was discarded do to its unacceptable fit indexes scores (CFI = 0.75, RMSEA = 0.038, and SRMR = 0.074 for perpetration; CFI = 0.91, RMSEA = 0.023, and SRMR = 0.051, for victimization).

Given the correlations within-factor errors and similar content in the items (Hooper et al., 2008), some modifications were made through the correlation of error terms to the four-factor model results (CFI = 0.84, RMSEA = 0.030, and SRMR = 0.047 for perpetration; CFI = 0.88, RMSEA = 0.027, and SRMR = 0.05, for victimization). The error term correlations included for the perpetration model were: item 6 with 7, from the psychological aggression factor; and error term 12 with 14; and 15 with 13, from the mild physical violence. For the victimization model: correlation between error terms 12 and 14, and 13 with 9 from the mild physical aggression factor.

The criteria to include this correlations in the model was the strength of the modification indices (MI) and Expected Parameter Change (EPC) values for the residual covariance, as well as the obvious overlap of the item contents (Byrne, 2012). For example, correlation between error terms 12 and 14, was include in both models (perpetration and victimization) due to it had MI values of 28.97 and 23.92, respectively; and the evident similarity of items content: 12 "You have hit your boyfriend/girlfriend" and item 14 "You have slapped your boyfriend/girlfriend." Goodness-of-fit results of before (Model 1) and after the correlation of error terms (Model 2) that confirm the fit of the proposed models to the original version are presented in **Table 3**.

TABLE 2 | Cronbach's alpha and omega coefficients of the M-CTS subscales.

	Perpetration			Victimization		
	α	CI 95%	ω	α	CI 95%	ω
Argumentation	0.45	0.4–0.49	0.48	0.43	0.38–0.47	0.43
Psychological violence	0.65	0.62–0.67	0.64	0.66	0.64–0.69	0.67
Mild physical violence	0.77	0.75–0.78	0.80	0.78	0.77–0.80	0.81
Severe physical violence	0.71	0.69–0.73	0.73	0.43	0.39–0.47	0.44

α, Cronbach's Alpha coefficient; ω, omega coefficient.

TABLE 3 | Goodness-of-fit indexes used to assess confirmatory factor analysis for the M-CTS.

Model 1		
Index	**Perpetration**	**Victimization**
CFI	0.84	0.88
Number of free parameters	60	60
Root Mean Square Error Approximation (RMSEA)		
Estimate	0.030	0.027
Standardized Root Mean Square Residual (SRMR)		
Value	0.047	0.05
Model 2 (Including correlation of error terms)		
Index	**Perpetration**	**Victimization**
CFI	0.90	0.91
Number of free parameters	63	62
Root Mean Square Error Approximation (RMSEA)		
Estimate	0.024	0.024
Standardized Root Mean Square Residual (SRMR)		
Value	0.043	0.049

TABLE 4 | Standardize model results: STDYX Standardization of the M-CTS.

Item	Squared multiple correlations	Factor loading	Estimate/*SE*
Argumentation			
(1) ¿Tú has discutido de forma tranquila?	0.23	0.47	12.93***
(2) ¿Tú has buscado información para apoyar tu punto de vista?	0.42	0.65	16.20***
(3) ¿Tú has llamado o intentado llamar a otra persona para que te ayude a arreglar las cosas?	0.11	0.33	11.23***
Psychological violence			
(4) ¿Tú has insultado o maldecido a tu novio?	0.32	0.56	20.07***
(5) ¿Tú te has molestado al hablar de un tema y/o te has negado a hacerlo?	0.28	0.53	20.67***
(6) ¿Tú te has marchado molesto/a de la habitación de la casa o el lugar donde estaban discutiendo?	0.24	0.48	18.49***
(7) ¿Tú has llorado como consecuencia de una discusión?	0.19	0.44	15.43***
(8) ¿Tú has dicho o hecho algo para fastidiar o molestar a tu novio?	0.30	0.55	21.66***
Mild physical violence			
(9) ¿Tú has amenazado con golpear o lanzar algún objeto a tu novio/a?	0.31	0.55	11.99***
(10) ¿Tú has intentado sujetar físicamente a tu novio/a?	0.15	0.38	10.74***
(11) ¿Tú has lanzado algún objeto a tu no novio/a?	0.41	0.64	15.82***
(12) ¿Tú has golpeado a tu novio/a?	0.41	0.64	13.25***
(13) ¿Tú has empujado o agarrado a tu novio/a?	0.52	0.72	20.09***
(14) ¿Tú le has dado una cachetada a tu novio/a?	0.34	0.68	11.08***
(15) ¿Tú has pateado o mordido a tu novio/a?	0.31	0.56	14.68***
Severe physical violence			
(16) ¿Tú has intentado ahogar a tu novio/a?	0.23	0.48	2.26*
(17) ¿Tú has dado una paliza a tu novio/a?	0.55	0.74	8.52***
(18) ¿Tú has amenazado a tu novio con un cuchillo o algún arma?	0.67	0.82	5.17***

Perpetration subscale.
****Two-tailed p-value < 0.001; *Two-tailed p-value < 0.05.*

TABLE 5 | Standardize model results: STDYX standardization of the M-CTS.

Item	Squared multiple correlations	Factor loading	Estimate/*SE*
Argumentation			
(1) ¿Tu novio/a ha discutido de forma tranquila?	0.12	0.34	9.48***
(2) ¿Tu novio/a ha buscado información para apoyar su punto de vista?	0.27	0.52	14.75***
(3) ¿Tu novio/a ha llamado o intentado llamar a otra persona para que ayude a arreglar las cosas?	0.22	0.47	14.32***
Psychological violence			
(4) ¿Tu novio/a te ha insultado o maldecido?	0.30	0.55	18.91***
(5) ¿Tu novio/a se ha molestado al hablar de un tema y/o se ha negado a hacerlo?	0.28	0.53	21.93***
(6) ¿Tu novio/a se ha marchado/molesto/o de la habitación de la casa o el lugar donde estaban discutiendo?	0.33	0.57	24.56***
(7) ¿Tu novio/a ha llorado como consecuencia de una discusión?	0.21	0.46	18.34***
(8) ¿Tu novio/a ha dicho o hecho algo para fastidiarte o molestarte?	0.33	0.58	25.02***
Mild physical violence			
(9) ¿Tu novio te ha amenazado con golpearte o lanzarte algún objeto?	0.43	0.66	16.52***
(10) ¿Tú novio ha intentado sujetarte físicamente?	0.24	0.49	14.64***
(11) ¿Tú novio te ha lanzado algún objeto?	0.33	0.57	11.62***
(12) ¿Tú novio te ha golpeado?	0.38	0.62	14.78***
(13) ¿Tu novio te ha empujado o agarrado?	0.61	0.78	27.80***
(14) ¿Tu novio te ha dado una cachetada?	0.29	0.53	9.9***
(15) ¿Tu novio te ha pateado o mordido?	0.40	0.64	17.74***
Severe physical violence			
(16) ¿Tu novio te ha intentado ahogar?	0.25	0.49	3.9***
(17) ¿Tu novio te ha dado una paliza?	0.23	0.48	3.98***
(18) ¿Tu novio te ha amenazado con un cuchillo o algún arma?	0.16	0.40	3.56***

Victimization subscale.
****Two-tailed p-value < 0.001.*

The final models obtained Good fit values in RMSEA and RSMR, and acceptable-fit values for CFI. It should be mention that the lack of convergence in the indexes values most not be understood as the model is misspecified or had any flaws in the data. It has been documented (Lai and Green, 2016) that this disagree arises because: (a) the two indexes by design, evaluate fit from different perspectives and, (b) the cut values of both are arbitrary and independent from each other.

Tables 4, **5** show the distribution of the items in each of the factors in perpetration and victimization models.

Known Groups Validity

Due to the distribution of the variables *Mann–Whitney U test* were performed in order to assess the ability of the M-CTS to contrasts of hypotheses of equality between means by sex and age. Along with the estimation of the statistical differences, the effect size was calculated though *A* static with Hanley y McNeil method, values around 0.010, 0.30, and 0.50, were considered as small, medium, and large, respectively. **Table 6** shows, as in previous studies (Fernández-Fuertes and Fuertes, 2010), significant statistical differences in scores between men and women. Higher levels of aggressiveness were self-reported by women in relation to men for the subscales of psychological violence ($Z = 7.91$; $p < 0.001$; $A = 0.39$) and mild physical violence ($Z = 4.59$; $p < 0.001$; $A = 0.52$). In the case of victimization, men self-reported significantly higher levels of victimization through psychological violence ($Z = 2.17$; $p < 0.05$; $A = 0.52$).

TABLE 6 | Means, standard deviations (SD), statistical differences and effect size by sex in the M-CTS subscales.

	Women (*N* = 1070; 57.5%)	Men (*N* = 791; 42.5%)	Total (*N* = 1861)	*Z*	*A*
Perpetration	*M (SD)*	*M (SD)*	*M (SD)*		
Argumentation	5.03 (2.58)	4.90 (2.51)	5.12 (2.62)	1.90	
Psychological violence	5.25 (2.68)	5.11 (2.61)	4.56 (3.47)	7.91***	0.39
Mild physical violence	5.13 (3.66)	4.37 (3.24)	1.09 (2.30)	4.59***	0.44
Severe physical violence	3.79 (3.02)	4.81 (3.63)	0.02 (0.27)	1.65	
Victimization					
Argumentation	1.32 (2.59)	1.03 (2.20)	4.99 (2.56)	1.81	
Psychological violence	0.79 (1.78)	1.29 (2.75)	4.56 (3.42)	2.17*	0.52
Mild physical violence	0.03 (0.32)	0.03 (0.30)	1.14 (2.45)	1.64	
Severe physical violence	0.02 (0.18)	0.04 (0.30)	0.03 (0.30)	0.25	

**Two-tailed p-value p < 0.05; ***Two-tailed p-value p < 0.001.*
A values around 0.010, 0.30, and 0.50, were considered as small, medium, and large, respectively.

To analyze the differences by age, the participants were grouped into early adolescence (12–14 years) and late adolescence (15–18 years) according to the criteria on the physical and mental development of the adolescents proposed by the United Nations International Children's Emergency Fund (UNICEF, 2011). Consistent with previous studies' findings (Foshee et al., 2009), the violent behaviors were self-reported

TABLE 7 | Means, SD and differences by age in the M-CTS subscales.

	12–14 years	15–18 years	Total		
	(*n* = 370; 19.9%)	(*n* = 1491; 80.1%)	(*n* = 1861)	*Z*	*A*
Perpetration	*M (SD)*	*M (SD)*	*M (SD)*		
Argumentation	4.75 (2.59)	5.21 (2.62)	5.12 (2.62)	2.92**	0.55
Psychological violence	4.05 (3.33)	4.68 (3.49)	4.56 (3.47)	3.22***	0.55
Mild physical violence	1.10 (2.42)	1.09 (2.27)	1.09 (2.30)	0.52	
Severe physical violence	0.03 (0.28)	0.02 (0.27)	0.02 (0.27)	0.28	
Victimization	*M (SD)*	*M (SD)*	*M (SD)*		
Argumentation	4.73 (2.66)	5.05 (2.53)	4.99 (2.56)	2.16*	0.54
Psychological violence	5.05 (2.53)	4.70 (3.44)	4.56 (3.42)	3.85***	0.56
Mild physical violence	1.17 (2.55)	1.13 (2.43)	1.14 (2.45)	0.17	
Severe physical violence	0.02 (0.29)	0.04 (0.30)	0.03 (0.30)	0.91	

Two-tailed p-value p < 0.05;*Two-tailed p-value p < 0.001.*

A values around 0.010, 0.30, and 0.50, were considered as small, medium, and large, respectively.

TABLE 8 | Spearman correlations between the M-CTS and DJTS and AQ scales, Means, SD.

			Mild	Severe		
		Psychological	physical	physical		
	Argumentation	violence	violence	violence	*M*	*SD*
Perpetration						
Argumentation	–				5.12	2.62
Psychological violence	0.27***	–			4.56	3.47
Mild physical violence	0.07**	0.42***	–		1.09	2.30
Severe physical violence	−0.03	0.05*	0.17***	–	0.02	0.27
Dominating Tactics	0.13***	0.42***	0.31***	0.097***	1.10	2.04
Jealous Tactics	0.22***	0.47***	0.22***	0.04*	3.04	2.85
AQ-verbal aggression	0.056*	0.19***	0.16***	0.02	2.64	0.78
AQ-Physical aggression	−0.025	0.13***	0.16***	0.04*	2.35	0.78
Victimization						
Argumentation	–				4.99	2.56
Psychological violence	0.26***	–			4.56	3.42
Mild physical violence	0.09***	0.42***	–		1.14	2.45
Severe physical violence	−0.01	0.11***	0.21***	–	0.03	0.30
Dominating tactics	0.11***	0.45***	0.29***	0.12***	1.45	2.53
Jealous tactics	0.23***	0.48***	0.20***	0.06*	3.59	3.26

Perpetration and Victimization.

**Two-tailed p-value < 0.05; **Two-tailed p-value < 0.01; ***Two-tailed p-value < 0.001.*

more frequently by the group of late adolescents. **Table 7** shows significant differences in the scales of perpetration in argumentation ($Z = 2.92$; $p < 0.005$; $A = 0.55$) and psychological violence ($Z = 3.22$; $p < 0.001$; $A = 0.55$).

Differences in the victimization self-reported aggressions are also shown in **Table 7**, there were significant differences for the subscales of argumentation ($Z = 2.16$; $p < 0.05$; $A = 0.54$) which had higher prevalences in late adolescents, and in the psychological violence which had higher prevalences in early adolescents ($Z = 3.85$; $p < 0.001$; $A = 0.56$).

Convergent Validity

Finally, Spearman correlations were calculated between M-CTS subscales, and for the scores of physical aggression and verbal aggression of the AQ scale with the perpetration subscales of the M-CTS, as well as the correlations between the DJTS subscales and the perpetration and victimization subscales of the M-CTS (**Table 8**). As expected, all correlations were statistically significant, except five; four of them from the perpetration subscales: (a) argumentation and severe physical violence, (b) argumentation and physical aggression of the AQ, (c) severe physical violence and verbal aggression subscale of AQ, and (d) severe physical violence and Jealous Tactics from DJTS. And, one from the victimization subscales (e) argumentation and severe physical violence from the M-CTS.

Administration Guidelines

The following specifications are recommended to administrate the test. First, researchers should inform the participants about the objectives and purposes of the study. Second, the researchers must obtain the informed consent of the adolescents, parents or legal guardians, and school's principals. Also, it is important that the researcher maintain the anonymity of participants' responses to the test. The researcher should read the test instructions in groups and explain the answer format with an example (first item) and should resolve participants' doubts before starting the test administration. Next, the results should be scored by two or three evaluators trained by experts per group. Finally, the researcher should allow 50 min for the test administration.

Score Scales and Interpretation Guidelines

Once the reliability and validity of the M-CTS in Mexican adolescents were tested and found acceptable, the Mexican scale's properties were qualitatively compared with those obtained by the Spanish version to identify the equivalence of the construct and factor structure consistence, in both populations. In both the Mexican and Spanish versions of the scale, the model of equations calculated through the confirmatory factor analysis obtained satisfactory scores in RMSEA, and CFI; this outcome verified the structural and functional statistics qualities of the scale in both populations.

DISCUSSION

The incorporation of the methodology proposed by the ITC to adapt the M-CTS for Mexican adolescents represents a remarkable advance for the dating violence research field in México. It makes possible—by contemplating cultural and linguistic variables of the nation—the consensual, rigorous, and reliable measurement of the problem. The results provide an indispensable base for the development of effective intervention and prevention programs (Borsa et al., 2012).

This adaptation represents, in addition, an improvement to the previous analysis of the M-CTS in the Spanish population; in the present study, in addition to a confirmatory factor analysis, known groups and concurrent validity analyses were conducted. These improvements provide greater evidence of the adequate psychometric guarantees and abilities of the M-CTS to respond to the current measurement demands of dating violence (Straus, 2004).

Nevertheless, it should be noted that as topic of future investigations, it would be interesting to test the measurement invariance of the M-CTS to ensure suitable group comparisons between men and women, or in between group ages. We strongly recommend to implement specific statistical procedures to test Differential Item Functions based on Classical Test Theory as Logistic Regressions or Lord Chi-square calculation based on the Item Response Theory, for example (Çokluk et al., 2016).

It is important to mention that the six modified items in this version proved to have adequate psychometric properties for measuring dating violence in Mexico because they obtained in each case a factorial weight above 0.40. The total scale and subscales obtained acceptable levels of reliability and validity and also demonstrated an equal factor structure to the one proposed in the literature and the previous validation studies (Fernández-González et al., 2013). These results position the M-CTS as one of the best scales for cross-cultural studies of dating violence.

After carrying out the adaptation, the usefulness of the methodology proposed by International Test Commission [ITC] (2016) was confirmed, as was the need for internationally recognized guides for the development and adaptation of scales. Otherwise, by continuing the use of the scales without carrying out the necessary adaptations—through proven and agreed procedures—for the populations of interest, there will be a great risk of reporting data that, instead of reflecting the problem, will report deficiencies in the scales, differences in the factorial structure, or measurement variances (Eremenco et al., 2005; Gjersing et al., 2010).

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Ethical approval for all procedures involving human subjects and analyses conducted for the current manuscript was provided by the Research Ethics Committee of the Autonomous University of Madrid (CEI-85-1576) in accordance with federal regulations governing human subjects research and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants, their parents, and school's supervisors and principals.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Andreu, J. M., Peña, M. E., and Graña, J. L. (2002). Adaptación psicométrica de la versión española del Cuestionario de Agresión [Psychometric adaptation of the Spanish version of the aggression questionnaire]. *Psicothema* 14, 476–482.

Boira, S., Chilet-Rosell, E., Jaramillo-Quiroz, S., and Reinoso, J. (2017). Sexism, distorted thoughts and violence in couple relationships in ecuadorian universities with students related to welfare and health. *Univers. Psychol.* 16, 12–24.

Borsa, J. C., Damásio, B. F., and Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: some consideratios. *Paidéia* 22, 423–432. doi: 10.1590/1982-43272253201314

Buss, A. H., and Perry, M. (1992). The aggression Questionnarie. *J. Pers. Soc. Psychol.* 63, 452–459. doi: 10.1037/0022-3514.63.3.452

Byrne, B. M. (2012). *Structural Equation Modeling With Mplus: Basic Concepts, Aplications and Programming.* New York, NY: Routledge.

Byrne, B. M., and Van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: addressing the issue of nonequivalence. *Int. J. Test.* 10, 107–132. doi: 10.1080/15305051003637306

Carrillo-Flores, I. (2014). Acciones para prevenir la violencia en mujeres en educación superior. *Caleidoscopio* 31, 157–172. doi: 10.33064/31crscsh540

Cascardi, M., and Avery-Leaf, S. (2015). Gender differences in dating aggression and victimization among low-income, urban middle school students. *Partner Abuse* 6, 383–402. doi: 10.1891/1946-6560.6.4.383

Cascardi, M., Avery-leaf, S., O'Leary, D., and Smith, A. M. (1999). Factor structure and convergent validity of the Conflicts tactics scale in high school students. *Psychol. Assess.* 11, 546–555. doi: 10.1037/1040-3590.11.4.546

Caulfield, M., and Riggs, D. (1992). The assessment of dating aggression: empirical evaluation of the Conflict tactics scale. *J. Inteterpers. Violence* 7, 549–558. doi: 10.1177/088626092007004010

Celis-Sauce, A., and Rojas-Solís, J. L. (2015). Adolescentes mexicanos como víctimas y perpetradores de violencia en el noviazgo [Mexican Adolescents as Victims and Perpetrators of Dating Violence]. *Reidocrea* 4, 60–65.

Chaín-Pinzón, N., Lorenzo-Seva, U., and Vigil-Colet, A. (2012). Psychometric characteristics of the Colombian adaptation of the aggression questionnaire of buss and perry in a sample of preadolescents and adolescents of Bucaramanga. *Univers. Psichol.* 3, 979–988.

Çokluk, Ö, Gül, E., and Dogan-Cül, C. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educ. Sci. Theory Pract.* 16, 319–330. doi: 10.12738/estp.2016.1.0329

Crutzen, R., and Ygram, G. J. (2017). Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all. *Health Psychol. Rev.* 11, 242–247. doi: 10.1080/17437199.2015.1124240

Domínguez-Lara, S. E., and Merino-Soto, C. (2015). ¿Por qué es importante reportar los intervalos de confianza del coeficiente alfa de Cronbach. *Rev. Latino. Cien. Social. Nilez Juventud* 13, 1326–1328. doi: 10.1080/17437199.2015.1124240

Eremenco, S. L., Cella, D., and Arnold, B. J. (2005). A comprehensive method for the translation and crosscultural validation of health status questionnaires. *Eval. Health Prof.* 28, 212–232. doi: 10.1177/0163278705275342

Fernández-Fuentes, A., Orgáz, M. B., and Fuentes, A. (2011). Características del comportamiento agresivo en las parejas de los adolescentes españoles. *Psicol. Conductual* 19, 501–522. doi: 10.1177/0163278705275342

Fernández-Fuertes, A. A., and Fuertes, A. (2010). Physical and psychological aggression in dating relationships of Spanish adolescents: motives and consequences. *Child Abuse Neglect* 34, 183–191. doi: 10.1016/j.chiabu.2010.01.002

Fernández-González, L., O'Leary, K. D., and Muñoz-Rivas, M. J. (2013). We are not joking: need for controls in reports of dating violence. *J. Interpers. Violence* 28, 602–620. doi: 10.1177/0886260512455518

Foshee, V. A., Benefield, T., Suchindran, C., Ennett, S. T., Bauman, K. E., Karriker-Jaffe, K. J., et al. (2009). The development of four types of adolescent dating abuse and selected demographic correlates. *J. Res. Adolesc.* 19, 380–400. doi: 10.1177/0886260512455518

Gjersing, L., Caplehorn, J. R. M., and Clausen, T. (2010). Cross-cultural adaptation of research instruments: language, setting, time and statistical considerations. *BMC Med. Res. Methodol.* 10:13. doi: 10.1186/1471-2288-10-13

Hambleton, R. K., and Patsula, L. (1999). Increasing the validity of adapted tests: myths to be avoided and guidelines for improving test adaptation practices. *J. Appl. Test. Technol.* 1, 1–12. doi: 10.1186/1471-2288-10-13

Hokoda, A., Ramos-Lira, L., Celaya, P., Vilhauer, H., Angeles, M., Ruíz, S., et al. (2006). Reliability of translated measures assessing dating violence among mexican adolescents. *Violence Vict.* 21, 117–127. doi: 10.1891/088667006780927367

Hooper, D., Coughlan, J., and Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *Electron. J. Bus. Res. Methods* 6, 53–60. doi: 10.1016/j.acap.2015.07.001

IBM (2011). *IBM SPSS Statistics for Windows, Version 20.0*. Armonk, NY: IBM Corp. doi: 10.1016/j.acap.2015.07.001

International Test Commission [ITC] (2005). *ITC Guidelines for Translating and Adapting Tests Versión 1.0*. Available at: www.InTestCom.org

International Test Commission [ITC] (2016). *The ITC Guidelines for Translating and Adapting Tests*, 2nd Edn. Available at www.InTestCom.org

Jöreskog, K. G. (2001). *Analysis of Ordinal Variables 2: Cross-Sectional Data. Text of the Workshop Estructural Equiation Modeling With LISREL 8.51*. Jena: Friederich-Shieller-Universitat Jena.

Kasian, M., and Painter, S. L. (1992). Frequency and severity of psychological abuse in a dating population. *J. Interpers. Violence* 7, 350–364. doi: 10.1177/088626092007003005

Lai, K., and Green, S. B. (2016). The problem with having two watches: assessment of fit when RMSEA and CFI disagree. *Multivariate Behav. Res.* 51, 220–239. doi: 10.1080/00273171.2015.1134306

López-Cepero, B., Rodríguez-Franco, J., and Rodríguez-Díaz, F. (2015). Measuring intimate partner abuse: a review of behavioral assessment tools. *Revista Iberoamericana de Diagnóstico y Evaluación - e Avaliação Psicológica* 2, 37–50.

Muñiz, J., Elosua, P., and Hambleton, R. K. (2013). International test commission guidelines for test translation and adaptation second edition: background. *Psicothema* 25, 151–157. doi: 10.7334/psicothema2013.24

Muñiz, J., Hernández, A., and Ponsada, V. (2015). New guidelines for test use: research, quality control and security of tests. *Papeles del Psicólogo* 36, 161–173.

Muñoz-Rivas, M. J., Andreu, J. M., Graña, J. L., O'Leary, K. D., and González, M. P. (2007a). Validation of the modified version of the Conflicts tactics scale (M-CTS) in a Spanish population of youths. *Psicothema* 19, 693–698.

Muñoz-Rivas, M. J., Graña, J. L., O'Leary, K. D., and González, M. P. (2007b). Physical and psychological aggression in dating relationships in Spanish university students. *Psicothema* 19, 102–107.

Muñoz-Rivas, M. J., Graña, J. L., O'Leary, K. D., and González, M. P. (2009). Prevalence and predictors of sexual aggression in dating relationships of adolescents and young adults. *Psicothema* 21, 234–240.

Muñoz-Rivas, M. J., Redondo, N., Zamarrón, M. D., and González, M. P. (2019). Violence in dating relationships: validation of the dominating and Jelous Tactis Scale in Sapnish youth. *Anales de Psicología* 35, 11–18. doi: 10.6018/analesps.35.1.319251

Muthén, L. M. P., and Muthén, B. (1998–2015). *M-plus User's Guide (Version 7)*. Los Angeles, CA: Muthén & Muthén.

Neidig, P. M. (1986). *The Modified Conflict Tactics Scale*. Beaufort, SC: Behavioral Sciences Associates.

Nocentini, A., Menesini, E., Pastorelli, C., Conolly, J., Pepler, D., and Craig, W. (2011). Physical dating aggresion in adolescece. *Eur. Psychol.* 16, 278–287. doi: 10.1027/1016-9040/a000045

Oakland, T., Poortinga, Y. H., Schlegel, J., and Hambleton, R. K. (2009). Directions international test commission: its history, current status, and future directions. *Int. J. Test.* 1, 3–32. doi: 10.1207/S15327574IJT0101

Oliva-Zárate, L., Rivera-Vargas, E. A., González-Flores, M. P., and Yedra, L. R. (2018). Violencia en el noviazgo en adolescentes de Veracruz, México. *Psique* 14, 8–24. doi: 10.26619/2183-4806.XIV.1.1

Pan, H., Neidig, P., and O'Leary, D. (1994). Male-female and agresor-victim differences in the factor structure of the modified conflicto tactics scale. *J. Interpers. Violence* 9, 366–382. doi: 10.1177/088626094009003006

Peña-Cárdenas, F., Zamorano-González, B., Hernández-Rodríguez, G., Hernández-González, M. L., Vargas-Martínez, J. I., and Parra-Sierra, V. (2013). Violencia en el noviazgo en una muestra de jóvenes mexicanos. *Revista Costarricense de Psicología* 32, 25–40.

Reichenheim, M., and Moraes, C. (2007). Operationalizing the cross-cultutral adaptation of epidemological measurement instruments. *Revista de Saúde Publica* 41, 1–10.

Rey-Anacona, C. A., González Cruz, Y. C., Sánchez Jiménez, V., and Saavedra, E. (2017). Sexism and dating violence in Spanish, Chilean and Colombian adolescents. *Behav. Psychol.* 25, 297–314.

Rodríguez, J. A. (2014). Violencia en el noviazgo de estudiantes universitarios venezolanos [Dating violence in Venezuelan University Students]. *Archivos de Criminología, Seguridad Privada y Criminalística* 2, 1–20.

Rojas-Solís, J. L., Fuertes-Martin, J. A., and Orgaz-Baz, M. B. (2017). Dating violence in young Mexican couples: a dyadic analysis. *Int. J. Soc. Psychol.* 32, 566–596.

Rubio-Garay, F., López-González, M. A., Carrasco, M. A., and Amor, P. J. (2017). Prevalencia de la violencia en el noviazgo: una revisión sistemática. *Papeles del Psicólogo* 38, 135–147. doi: 10.23923/pap.psicol2017.2831

Rubio-Garay, F., López-González, M. A., Saúl, L. A., and Sánchez-Elvira-Paniagua, A. (2012). Direccionalidad y expresión de la violencia en las relaciones de noviazgo de los jóvenes. *Acción Psicológica* 9, 61–70. doi: 10.5944/ap.9.1.437

Ryan, A. M., Chan, D., Ployhart, R. E., and Slade, L. A. (1999). Employee attitude surveys in a multinational organization: considering language and culture in assessing measurement equivalence. *Pers. Psychol.* 52, 37–58. doi: 10.1111/j.1744-6570.1999.tb01812.x

Ryan, K. (2013). Issues of reliability in measuring intimate partner violence during courtship. *Sex Roles* 69, 131–148. doi: 10.1007/s11199-012-0233-4

Shaffer, C. M., Corona, R., Sullivan, T. N., Fuentes, V., and McDonald, S. E. (2018). Barriers and supports to dating violence communication between latina adolescents and their mothers: a qualitative analysis. *J. Fam. Violence* 33, 133–145. doi: 10.1007/s10896-017-9936-1

Straus, M. A. (2004). Cross-cultural reliability and validity of the revised Conflict tactics scales: a study of university student dating couples in 17 nations. *Cross Cult. Res.* 38, 407–432. doi: 10.1177/1069397104269543

UNICEF (2011). *The State of the World's Children 2011*. New York, NY: United Nations children's fund.

Vega-Valero, C. Z. (2015). Las estrategias de afrontamiento ante la violencia en el noviazgo. *Revista Digital Internacional de Psicología y Ciencia Social* 1, 133–140. doi: 10.22402/rdipycs.unam.1.1.2015.31.133-140

Ventura-León, J. L., and Caycho-Rodríguez, T. (2017). El coeficiente Omega: un método para la estimación de la confiabilidad. *Rev. Latino. Cien. Social. Niñez Juventud* 15, 625–627.

World Health Organization [WHO] (2002). *World Report on Violence and Heatlh*. Washington, DC: World Health Organization.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 16 April 2019
doi: 10.3389/fpsyg.2019.00861

An Evaluation of the Belief in Science Scale

Neil Dagnall, Andrew Denovan, Kenneth Graham Drinkwater and Andrew Parker*

Department of Psychology, Manchester Metropolitan University, Manchester, United Kingdom

The Belief in Science Scale (BISS) is a unidimensional measure that assesses the degree to which science is valued as a source of superior knowledge. Due to increased academic interest in the concept of belief in science, the BISS has emerged as an important measurement instrument. Noting an absence of validation evidence, the present paper, via two studies, evaluated the scale's factorial structure. Both studies drew on data collected from previous research. Study 1 (N = 686), using parallel analysis and exploratory factor analysis, identified a unidimensional solution accounting for 56.43% of the observed variance. Study 2 (N = 535), using an independent sample, tested the unidimensional solution using confirmatory factor analysis (CFA). Data-model fit was good (marginal for RMSEA): CFI = 0.93, TLI = 0.91, RMSEA = 0.09 (90% CI of 0.08 to 0.10), SRMR = 0.04. Invariance testing across gender supported invariance of form, factor structure, and item intercepts for this one-factor model. BISS at the overall level correlated negatively with the reality testing dimension of the Inventory of Personality Organization (IPO-RT), demonstrating convergent validity. Researchers often use the IPO-RT as an indirect index of preference for experiential processing (intuitive thinking). In this context, only BISS scores above the median (second quartile) produced a reduction in experiential-based thinking. The authors discuss these findings in the context of belief in science as a psychometric construct.

Keywords: belief in science, psychometric validation, reality testing, thinking style, convergent validity

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Silvia Testa,
University of Turin, Italy
Hugo Carretero-Dios,
University of Granada, Spain

****Correspondence:***
Neil Dagnall
n.dagnall@mmu.ac.uk

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *10 December 2018*
Accepted: *01 April 2019*
Published: *16 April 2019*

Citation:
Dagnall N, Denovan A, Drinkwater KG and Parker A (2019) An Evaluation of the Belief in Science Scale. Front. Psychol. 10:861. doi: 10.3389/fpsyg.2019.00861

INTRODUCTION

Beliefs are a fundamental aspect of human cognition that fulfill important individual and social functions. Explicitly, beliefs provide meaning, comfort, and communality (Hogg and Mulling, 1999; Heine et al., 2006). This is particularly true of religious faith, which is associated with a range of positive psychological benefits. These include moderating negative factors related to lack of control (Kay et al., 2009), reducing anxiety (Inzlicht et al., 2011) and decreasing stress (Ano and Vasconcelles, 2005). Farias et al. (2013) contend that secular beliefs, such as Humanism and political ideologies perform comparable functions within non-religious individuals (Gray, 2004).

Although science and religion offer competing, often contradictory explanations, at a deeper, conceptual level, research suggests that they perform comparable psychological functions (i.e., structure life, provide reassurance, and facilitate social integration) (Ziman, 1978/1991). In support of this notion, studies report that beliefs related to human advancement offer positive, compensatory psychological functions (Rutjens et al., 2009, 2010). Explicitly, higher levels of belief in science are associated with positive psychological outcomes, such as happiness, lower levels of stress and reduced death anxiety (Aghababaei et al., 2016).

Acknowledging the potentially important role that secular beliefs play in modern society, Farias et al. (2013) developed the Belief in Science Scale (BISS). The BISS is a 10-item research tool, which measures the degree to which individuals endorse the legitimacy of the scientific approach. Particularly, the BISS assesses belief in the value of science as an institution and a source of superior knowledge. Accordingly, the scale recognizes differences in attitudes toward science. These range from rejection of the scientific approach, through acceptance of science as a reliable but fallible source of knowledge, to the conviction that science provides exclusive, veridical insights into reality. The latter doctrinaire perspective depicts science as a unique, central value. Consistent with this, the defining features of belief in science are confidence and trust in the validity of scientific methods and outcomes. Furthermore, higher belief in science is associated with outright dismissal of notions that sit outside of the traditional scientific framework. This manifests typically as rejection of scientifically unsubstantiated beliefs (i.e., paranormal) and religious skepticism.

Farias et al. (2013) tested the notion that belief in science provides secular individuals with psychological meaning and comfort in threatening contexts by conducting two related studies. These necessitated the development of BISS. Prior to the first experiment, Farias et al. (2013) gave items assessing belief in science to a sample of 144 participants. Subsequent psychometric examination, in the form of exploratory factor analysis (varimax rotation), yielded a single dimension accounting for 57% of the variance. All items loaded (≥0.56) and the scale demonstrated high internal consistency (α = 0.86). The overall sample mean (M = 3.23, SD = 1.04) was consistent with moderate belief in science. In study two (N = 60), further consideration of the psychometric properties of BISS, also found good internal consistency (α = 0.88).

Following the initial evaluation, Farias et al. (2013) used the BISS in their experiments. The first, found that rowers in a high-stress condition (pre-completion) vs. low-stress condition (training) reported greater belief in science. This result was congruent with the notion that belief in science helps secular individuals cope with stress. Although, Farias et al. (2013) acknowledged that context manipulation (competition vs. training) might affect also scientific focus (i.e., encourage emphasis on training regimen and equipment).

Within the second experiment, participants were assigned randomly to one of two mortality salience conditions (thoughts and feeling about own death vs. experiencing dental pain; control) and completed self-report measures assessing scientific determinism (Paulhus and Carey, 2010), religiosity and affect (negative and positive) (Watson et al., 1988).

Noting potential construct overlap, a moderate positive correlation between belief in science and scientific determinism (Paulhus and Carey, 2010), Farias et al. (2013) conducted a principal components analysis (PCA) on all science-related items. This used oblimin rotation, an oblique solution that permits factor correlation. The PCA identified three related but distinct factors: belief in science, original 10-items (eigenvalue = 5.74, loadings ≥0.62); scientific determinism (environmental factors), 3-items (eigenvalue = 2.02, loadings ≥0.68); and scientific determinism (biological factors), 4-items (eigenvalue = 1.79, loadings ≥0.66). This outcome supported the supposition that belief in science, although correlated with scientific determinism, was a separate construct. Consistent with study one outcomes, analysis revealed that participants in the mortality salience condition (vs. controls) scored higher on belief in science.

Overall, findings were consistent with Farias et al.'s (2013) conceptualisation of science as a form of "faith" in secular individuals that facilitates coping in stressful and anxiety-provoking situations. Furthermore, Farias et al. (2013) concluded that analytical thinking, rational enquiry and consideration of empirical evidence were key characteristics associated with scientific thinking. In this context, belief in science places an emphasis on fact based, objective (vs. objective experiential) evidence.

The BISS has also demonstrated criterion validity across a range of studies. For instance, Irwin et al. (2016) reported a negative moderate correlation (r = −0.55) between belief in science and the New Age Beliefs subscale of the Survey of Scientifically Unsubstantiated Beliefs (SUBS) (Irwin and Marks, 2013). This was consistent with Irwin et al. (2015), who observed strong negative associations between BISS and SUBS subscales (New Age Beliefs, r = −0.63; Traditional Religious Beliefs, r = −0.71). Moreover, Irwin et al. (2015) reported a moderate negative relationship (r = −0.32) between BISS and The Inventory of Personality Organization (IPO–RT; Lenzenweger et al., 2001). The IPO–RT assesses self-reported proneness to deficits in reality testing and researchers often use the scale as an index of experiential, intuitive thinking style (Drinkwater et al., 2012; Dagnall et al., 2015a, 2018; Denovan et al., 2017b).

Consistent with this notion, Irwin et al. (2016) found that believers in the paranormal tended to discount the values of science, and preferred to endorse ideas based on their emotional (rather than their rational) appeal. Accordingly, believers subject decisions to less critical scrutiny. Irwin et al. (2016) concluded that these characteristics reflect opposing worldviews. The scientific perspective comprises presumptive skepticism and an acceptance of the values of science, whereas a subjective and anti-materialistic outlook on life typifies paranormal belief (Zusne and Jones, 1989). Generally, these findings concur with preceding work that indicates that faith in science, religion and the paranormal represent independent dimensions of belief (Williams et al., 1989; Ståhl et al., 2016).

Despite these encouraging outcomes, the BISS is psychometrically underdeveloped. Even though widely cited, researchers have yet to validate the BISS. Indeed, consideration of the literature reveals that other than the reported EFA, the BISS structure remains unsubstantiated. Furthermore, within studies employing the BISS, authors have either failed to include psychometric details (Valdesolo et al., 2016), or merely confirmed that the BISS possesses high internal consistency (i.e., Irwin et al., 2015, α = 0.93; Ståhl et al., 2016, α = 0.96). This lacks exactitude and rigor because scale analysis has failed to progress beyond EFA. Hence, further research is required to evaluate the measurement properties of the BISS.

Additionally, EFA is problematic when used in isolation because it merely identifies underlying factor structure within observed variables without reference to outcome (i.e., construct coherence). Typically, confirmatory factor analysis (CFA) is generally required to test the appropriateness of the emergent model (Suhr, 2006). This is consistent with psychometric theorists, who contend that scale development should start with exploration (EFA) then progress to CFA. CFA is preferable when measurement models possess a well-developed underlying theory for hypothesized patterns of loadings (Hurley et al., 1997). In the case of BISS, Farias et al. (2013) advocate a single, general factor underpinning belief in science. Hence, a thorough examination of scale structure is required in order to establish the conceptual constraints of the scale and determine its usefulness as a general measure of belief in science.

The present study examined the psychometric properties of the BISS by performing two related studies. Study 1 evaluated the analysis performed by Farias et al. (2013) via utilizing Horn's parallel analysis in addition to EFA. This was necessary to examine the replicability of Farias et al.'s (2013) results in an EFA context. Study 2 comprised a test of the resultant factor model from study 1 using CFA. Invariance testing followed an analysis of general factor structure, by assessing the degree to which different groups (males and females) performed on the measure. Invariance testing provides a further level of psychometric scrutiny by evaluating the extent to which scores reflect true differences across groups as opposed to artifacts of measurement bias (Brown, 2006; Byrne, 2010; Denovan et al., 2017a). Study 2 extended the preceding study by testing the emergent factor structure within an independent sample, and by assessing the convergent validity of BISS. Convergent validity is useful to assess whether a measure of a specific construct aligns with another measure it should theoretically relate to. The IPO-RT was an appropriate measure because it is a known correlate of belief in science, which indexes intuitive thinking. Specifically, the IPO-RT assesses proneness to reality testing deficits (Dagnall et al., 2014, 2015b, 2018). Explicitly, "the capacity to differentiate self from non-self, intrapsychic from external stimuli, and to maintain empathy with ordinary social criteria of reality" (Kernberg, 1996, p. 120). This delineation is consistent with Langdon and Coltheart's (2000) information-processing style account of belief generation. Noting these conceptual features, researchers frequently use the IPO-RT as an index of experiential, intuitive thinking style (Drinkwater et al., 2012; Dagnall et al., 2015b; Denovan et al., 2017a).

MATERIALS AND METHODS

Data Collection and Procedure

In order to evaluate the psychometric properties of the BISS two independent samples of respondents were required. To create these, amalgamation of data sets from previously published studies and ongoing research projects was undertaken. The researchers collected all data via online survey. In total, this comprised five merged data sets. Researchers have previously successfully utilized this method to generate large heterogeneous samples. Prominent examples are Revised Paranormal Belief Scale (Drinkwater et al., 2017), and Australian Sheep Goat Scale (Drinkwater et al., 2018).

Integration of BISS data sets was apposite since the research team have previously used the measure in comparable self-report studies. These have addressed a range of diverse research questions. The main advantage of data merging is the generation of sample sizes that permit the use of sophisticated statistical techniques. Explicitly, the combining data increases sample size, enhances statistical power and produces greater within sample variation (Van der Steen et al., 2008). This is particularly important when using procedures such as CFA, which require as many cases as possible (Brown, 2006). Hence, consolidation of BISS data was a convenient method that utilized existing, previously screened data to meet analytical constraints. Moreover, this approach generates a sample that would be difficult to recruit because of cost and time limitations.

Data collection for both studies occurred between September 2012 and September 2016 (see section "Ethics"). Recruitment was by emails to students (undergraduate and postgraduate) enrolled on healthcare programs (Nursing, Physiotherapy, Psychology, Speech, and Language Therapy, etc.), staff across faculties at the Manchester Metropolitan University, and local businesses/community groups. There were two exclusion criteria. Firstly, respondents had to be at least 18 years of age. Secondly, in order to prevent multiple responses instructions stated that respondents must not participate if they had undertaken similar or related research.

In all cases, respondents within the original research completed the BISS alongside several other measures. These assessed cognitive-perceptual personality factors, decision-making and anomalous beliefs (i.e., Irwin et al., 2015, 2016). In study 1, the BISS did not appear alongside the IPO-RT, whereas study 2 data derived from instances where the BISS and IPO-RT appeared within the same set of measures.

All studies employed the same, routine standardized procedures. Before undertaking the measures potential respondents received detailed information from the researchers. This outlined study aims, purpose, content, and ethical procedures. Assenting respondents provided informed consent via a survey option confirming willingness to participate. Subsequently, respondents received the study materials. Together with study measures there was a brief demographic section requesting age, preferred gender, and course of study if student, or occupation. Procedural instructions were consistent across studies. They directed respondents to progress through sections systematically, respond to items in an open and honest manner, work at their own pace, and reassured respondents that there were no right or wrong answers. To prevent potential order effects section order rotated across respondents.

Ethics Statement

The research team gained ethical authorization for a program of studies exploring relationships between anomalous beliefs, decision-making and cognitive-perceptual personality factors as part of the grant bidding process. In total, there were three bi-annual calls (September 2012, 2014, and 2016). Review rated each application as routine and granted ethical approval. The Director of the Research Institute for Health and Social Change (Faculty of Health, Psychology and Social Care) and Ethics Committee within the Manchester Metropolitan University supervised this process. This process demanded that two experienced reviewers scrutinized the documentation. If research, as in this case, is classified as routine this constitutes full ethical approval. This was the required level of institutional approval at that point in time.

Respondents

Study 1

The data set for study 1 contained 686 respondents. The mean (M) sample age was 26.70 years (SD = 11.07, range = 18–69 years). Disaggregation by gender revealed that 279 (40%) respondents were male and 407 (60%) female. Skewness and kurtosis values were within the recommended range of −2.0 to +2.0 (Byrne, 2010; **Table 1**). However, examination of multivariate normality suggested non-normality, as Mardia's (1970) skewness ($b1p$ = 9.80, $p < 0.001$) and kurtosis estimates ($b2p$ = 29.737, $p < 0.001$) indicated significant deviation from a normal distribution.

Study 2

The Study 2 sample comprised 534 (262, 49% male; 272, 51% female) respondents who had completed both the BISS and the IPO-RT. Mean (M) sample age was 37 (SD = 14.74, range = 18–71 years). All items, with the exception of IPO-RT items 4 and 16, demonstrated acceptable univariate skewness and kurtosis (i.e., between −2.0 and +2.0) (**Table 1**). Although, multivariate non-normality existed (skewness: $b1p$ = 130.27, $p < 0.001$; kurtosis: $b2p$ = 52.28, $p < 0.001$).

Measures

Study 1

The only measure examined in Study 1 was the BISS. The BISS is a 10-item, self-report tool that assesses level of epistemic beliefs related to science. Specifically, items reference notions of scientific pre-eminence (i.e., the idea that science possesses unique and central value that provide a superior, exclusive guide to reality) (Farias et al., 2013; Valdesolo et al., 2016). Items take the form of statements (e.g., "We can only rationally believe in what is scientifically provable"), and respondents indicate level of agreement via a 6-point Likert scale (ranging from 1 "strongly disagree" to 6 "strongly agree"). Thus, raw scores range from 10 to 60, with higher scores indicating stronger belief in science. Previous work reports that the BISS is unidimensional and possesses high internal consistency (Farias et al., 2013; Irwin et al., 2015).

Study 2

In study 2, alongside the BISS, respondents completed the IPO-RT subscale of The Inventory of Personality Organization (IPO–RT; Lenzenweger et al., 2001). Within the IPO-RT, there are 20-items presented as statements (e.g., "When everything around me is unsettled and confused, I feel that way inside"). Respondents indicate the degree to which they endorse each statement using a five-point Likert scale

TABLE 1 | Summary statistics for all Study 1 and Study 2 variables.

	Study 1				Study 2								
BISS item	***M***	***SD***	**Skew.**	**Kurt.**	***M***	***SD***	**Skew.**	**Kurt.**	**IPO-RT item**	***M***	***SD***	**Skew.**	**Kurt.**
Q1	4.90	1.41	−1.38	1.09	5.21	1.27	−1.67	1.98	Q1	2.98	0.88	−0.19	0.04
Q2	4.05	1.46	−0.55	−0.48	4.73	1.44	−1.04	0.19	Q2	2.19	1.01	0.35	−0.79
Q3	3.97	1.53	−0.49	−0.82	4.27	1.67	−0.66	−0.85	Q3	2.47	0.97	0.22	−0.53
Q4	3.36	1.60	−0.03	−1.23	3.52	1.74	−0.17	−1.35	Q4	1.41	0.70	1.90	3.91
Q5	3.31	1.53	0.01	−1.09	3.65	1.69	−0.14	−1.25	Q5	2.43	0.97	0.15	−0.44
Q6	3.54	1.59	−0.15	−1.14	4.01	1.79	−0.47	−1.14	Q6	2.22	1.05	0.59	−0.32
Q7	3.19	1.60	0.17	−1.14	3.52	1.78	−0.08	−1.37	Q7	1.66	0.92	1.29	1.00
Q8	3.49	1.56	−0.12	−1.05	3.84	1.63	−0.37	−1.02	Q8	1.51	0.84	1.59	1.80
Q9	4.09	1.49	−0.65	−0.44	4.46	1.59	−0.89	−0.29	Q9	2.05	1.10	0.66	−0.64
Q10	4.57	1.32	−0.84	0.26	5.02	1.23	−1.18	0.73	Q10	2.14	0.96	0.48	−0.38
									Q11	1.71	0.85	1.07	0.68
									Q12	1.63	0.89	1.20	0.49
									Q13	1.93	0.91	0.85	0.46
									Q14	2.02	1.02	0.77	−0.18
									Q15	2.02	1.16	0.90	−0.18
									Q16	1.54	0.97	1.89	2.94
									Q17	2.26	1.07	0.29	−0.88
									Q18	1.71	0.88	1.12	0.63
									Q19	1.61	0.87	1.25	0.67
									Q20	2.13	1.01	0.68	−0.02

BISS, Belief in Science Scale, IPO-RT, Inventory of Personality Organization-Reality Testing subscale.

(1 = never true to 5 = always true). Accordingly, total scores range from 20 to 100, with higher scores reflecting subjective evaluation of perceived likelihood of reality testing errors. Researchers often use IPO-RT scores as an index of intuitive thinking style (Denovan et al., 2017b). This derives from the supposition that the IPO-RT references suspension of reality testing, external critical evaluation (Irwin, 2004). Studies have established the psychometric properties of the IPO-RT. Particularly the measure possesses construct validity and demonstrates excellent internal consistency ($\alpha = 0.90$; $\omega = 0.93$) and test–retest reliability (Lenzenweger et al., 2001; Dagnall et al., 2018).

Data Analysis

Psychometric examination of the BISS progressed through a series of increasingly sophisticated analytical techniques. These included Horn's parallel analysis, exploratory factor analysis [EFA via maximum likelihood (MLR)], and CFA. The initial use of parallel analysis alongside scree plot assessment was necessary to judge the number of underlying factors. In addition, parallel analysis represents the most accurate approach to determine the quantity of factors to keep (Pallant, 2007). Accordingly, this included random resampling of the raw data (O'connor, 2000). EFA (SPSS 25) using the suggested number of factors then provided information on item loadings (Çokluk and Koçak, 2016).

Following parallel analysis and EFA, CFA conducted via Mplus 7.4 (Muthén and Muthén, 2015) assessed the appropriateness of data-model fit. Testing used the robust MLR method. This produces MLR parameter estimates and standard errors that are robust to instances of non-normality (Marsh et al., 2013).

The chi-square statistic (χ^2), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI) and absolute fit indices (Root-Mean-Square Error of Approximation, RMSEA; Standardized Root-Mean-Square Residual, SRMR) gaged model fit. The 90% confidence interval (CI) was included for RMSEA. CFI and TLI values >0.90 indicates good fit (Hopwood and Donnellan, 2010). According to Browne and Cudeck (1993), absolute values of 0.05, 0.06–0.08, and 0.08–1.0 reflect good, satisfactory, and marginal fit for RMSEA and SRMR.

Omega coefficient (estimated using JASP; Jeffreys's Amazing Statistics Program) determined internal consistency before invariance testing. This is a more effective reliability estimate than popular approaches such as coefficient alpha, which typically over- or underestimates the true reliability of a measure (Deng and Chan, 2017). Multi-group CFA examined invariance of factor structure (configural), factor loadings (metric), and item intercepts (scalar) in relation to gender for the superior factor solution. Chen's (2007) criteria of a CFI difference ≤ 0.01 and RMSEA ≤ 0.015 determined satisfactory fit for each invariance test.

In order to determine the replicability of the factor model from Study 1 in an independent sample, Study 2 analysis examined this model using CFA and measurement invariance. Also within Study 2, a test of convergent validity occurred. This involved comparing BISS with the criterion measure IPO-RT.

RESULTS

Study 1

For parallel analysis, eigenvalues from the raw data with values higher than those from the random data represent the resultant factors. A parallel analysis (with 1000 resamples) revealed that one factor (eigenvalue = 5.64) possessed an eigenvalue higher than random data (eigenvalue = 1.19). Therefore, one factor existed. Scree plot assessment further confirmed this. EFA examined the BISS with the restricted number of factors (Çokluk and Koçak, 2016). Results revealed satisfactory sampling adequacy; Kaiser-Meyer-Olkin measure (KMO) = 0.92 and a reasonable item correlation matrix, Bartlett's Test of Sphericity ($p < 0.001$). The single factor explained 56.43% of variance, and all factor loadings bar one (item 2) exceeded 0.4 (Norman and Streiner, 1994) with the majority of items (8 of 10) exceeding the strict factor loading requirements of 0.6 by Hair et al. (1998). Although item 2 loaded below 0.4, it exceeded the minimum cut-off of 0.32 suggested by Tabachnick and Fidell (2014). Lastly, examination of internal consistency revealed omega reliability was high for BISS, $\omega = 0.91$.

Study 2

A replication of the resultant one-factor model in study 1 with a separate dataset revealed (using CFA) good fit and marginal fit for RMSEA, χ^2 (35, $N = 534$) = 202.26, $p < 0.001$, CFI = 0.93, TLI = 0.91, RMSEA = 0.09 (90% CI of 0.08 to 0.10), SRMR = 0.04. Inspection of standardized parameter estimates (**Table 2**) reported a similar distribution of item loadings to study 1. Omega reliability was consistent with Study 1 (i.e., high for BISS, $\omega = 0.93$). In addition, for IPO-RT omega reliability was good, $\omega = 0.88$.

Multi-group analysis comparing gender revealed good model fit at the configural level across indices (excluding RMSEA), χ^2 (70, $N = 534$) = 239.73, $p < 0.001$, CFI = 0.93, TLI = 0.91, RMSEA = 0.09 (90% CI of 0.08 to 0.10), SRMR = 0.04. For metric invariance, an acceptable CFI difference of 0.005 existed alongside a minimal RMSEA difference of 0.002. Scalar invariance testing indicated a satisfactory difference for CFI (0.009) and RMSEA (0.001).

A test of convergent validity examined Pearson correlations between total BISS with Reality Testing (IPO-RT). Total BISS possessed a significant negative correlation with IPO-RT, $r(532) = -0.28$, $p < 0.001$ (95% CI of -0.36 to -0.19). *Post hoc* analyses split BISS at the quartile level to assess further its relationship with IPO-RT. A one-way ANOVA (using bootstrapping with 1000 resamples) indicated a differential relationship existed between BISS quartiles and IPO-RT, $F(3,530) = 17.62$, $p < 0.001$. Given the identification of non-normality in the data, bootstrapping enables a more accurate estimation of p-values and standard errors (Byrne, 2010). Indeed, bootstrapping performs well even in datasets of extreme non-normality (Nevitt and Hancock, 2001), and is a suitable alternative to MLR estimation considering an ANOVA command is not present in Mplus. The bootstrapping procedure generated estimations of standard errors alongside

TABLE 2 | Standardized parameter estimates for CFA in Study 2.

Item	Parameter estimate	R^2
Q1 Science provides us with a better understanding of the universe than does religion.	0.59**	0.35
Q2 "In a demon-haunted world, science is a candle in the dark." (Carl Sagan)	0.47**	0.22
Q3 We can only rationally believe in what is scientifically provable.	0.80**	0.65
Q4 Science tells us everything there is to know about what reality consists of.	0.79**	0.62
Q5 All the tasks human beings face are soluble by science.	0.82**	0.67
Q6 The scientific method is the only reliable path to knowledge.	0.90**	0.81
Q7 The only real kind of knowledge we can have is scientific knowledge.	0.88**	0.78
Q8 Science is the most valuable part of human culture.	0.76**	0.58
Q9 Science is the most efficient means of attaining truth.	0.81**	0.66
Q10 Scientists and science should be given more respect in modern society.	0.68**	0.46

**Indicate $p < 0.001$; all R^2-values statistically significant at $p < 0.001$.*

bias-corrected and accelerated CIs (at the 95% confidence level). Further scrutiny via mean comparisons tested the possibility that the relationship between BIS and IPO-RT was not linear. Using Bonferroni correction revealed, that whilst no differences were present between the first and second quartile, scores above the median differed significantly from those below the median. This indicates that a moderate level of BISS is required before a decline in intuitive thinking becomes evident (**Table 3**).

DISCUSSION

The present paper found that, consistent with Farias et al. (2013), a one-factor solution best explained BISS scores. Further psychometric consideration revealed that the measure demonstrated good/excellent internal consistency across the two studies (study 1, ω = 0.91, study 2, ω = 0.93). Examination of scale items indicated that respondents esteemed both the principles of science (i.e., providing meaning) and the application of science to specific applications (i.e., problem solving).

Studies 1 and 2 validated the one-factor solution, signifying that this was congruent with the single factor model advocated by Farias et al. (2013). Support for the one-factor solution was compelling because study 2 using an independent sample replicated the model tested in study 1. In terms of convergent validity, the BISS negatively correlated with reality testing ($r = -0.28$). The size of this relationship was similar to the correlation observed by Irwin et al. (2015) ($r = -0.32$). Overall, findings suggest that belief in science is moderately associated with the tendency to engage in experiential, intuitive thought. Within the present study, the BISS correlated negatively with the IPO-RT.

Collectively study findings indicated that higher levels of belief in science were associated with a lower propensity to reality testing deficits. A caveat to this statement was the observation that a decline in RT scores was evident only within participants scoring above the median on BISS, $r = -0.12$, $n = 269$, $p = 0.03$ (95% CI of −0.02 to −0.24). Below the median, there was no relationship between BISS and IPO-RT, $r = -0.01$, $n = 265$, $p = 0.449$ (95% CI of −0.14 to 0.13). This implies that moderate levels of BISS were required to facilitate a reduction in subjective, experiential-based thinking.

This view is consistent with the conceptual nature of scientific thinking. Explicitly, that analytical thinking is a key tenet of the scientific approach. This includes critical evaluation in the form of rational enquiry and objective consideration of evidence. These features are inherently contrary to intuitive thinking, which draws upon experiential, subjective appraisal of information. In this context, the findings are congruent with Farias et al.'s (2013) notion that higher levels of belief in science reflect a preference for analytical thinking. This typically manifests as a predilection for objective, external fact based (vs. subjective experiential) evidence.

Although these conclusions are congruent with previous research, there are limitations to consider. A particular concern is the size of the correlation between BISS and RT, which was only in the medium range. Indeed, the variables shared only approximately 7% variance. This is indicative of the fact that a range of factors in addition to belief in science influence thinking style. These include, but are not restricted to, motivation or ability to expend cognitive effort (Shiloh et al., 2002), and ability, in the form of task-relevant background knowledge or expertise (Novak and Hoffman, 2008). Accordingly, future studies should examine the degree to which these factors interact with belief in

TABLE 3 | Reality testing scores as a function of belief in science quartiles.

	Comparisons (mean differences) between quartiles	
Contrast	**Mean difference (*Sig.*)**	**95% BCa CI**
Quartile 1 vs. Quartile 2	0.74(0.537)	−1.58, 3.06
Quartile 1 vs. Quartile 3	4.99(< 0.001)**	2.49, 7.69
Quartile 1 vs. Quartile 4	7.75(< 0.001)**	5.39, 10.18
Quartile 2 vs. Quartile 3	4.25(0.004)*	1.71, 6.85
Quartile 2 vs. Quartile 4	7.01(< 0.001)**	4.60, 9.35
Quartile 3 vs. Quartile 4	2.76(0.026)*	0.48, 5.09

*Indicates $p < 0.05$, **indicates $p < 0.001$; 95% BCa CI: Bias-corrected and Accelerated confidence interval based on 1000 bootstrapped samples.*

science. It seems likely that high (vs. low cognitive) load and level of proficiency will influence the degree to which individuals appraise information, make decisions and draw on faith in science. With hindsight, the observation of a small correlation concurs with the view that the IPO-RT assesses a peculiar definition of thinking style. Specifically, one that indexes reality distortions and psychotic like phenomena (Lenzenweger et al., 2001).

A further concern is that both the BISS and IPO-RT are only "proxy" indirect measures of preferential thinking style. Accordingly, the scales do not directly assess thought. Instead, they index qualities reflective of the respective thinking style (Denovan et al., 2017b). In this context, it is important to note that BISS assesses "belief in the veracity of the scientific principles and methods," and IPO-RT taps the inclination to draw upon internal (rather than external) cognitions. Moreover, the present study failed to consider demographic factors such as level of education and occupational statues, which may indirectly influence critical thinking and belief in science. Thus, subsequent research could consider also the degree to which these factors affect belief in science.

Regarding BISS, there is an important distinction between confidence in the concept of science and the application of science based rationality. Many scientific informed discussions, such as those around climate change and the extinction of the dinosaurs, require systematic evaluation of information collected via methodical means. However, this process is often truncated, or terminated prematurely. This is often the case when individuals hold strong views about a topic and select (either consciously or unconsciously) evidence that supports their perspective. This assimilation bias leads to the dismissal of disconfirming evidence (Lord et al., 1979; Whitmarsh, 2011). Hence, it is possible to have a high belief in science, but base decision making on experiential (intuitive) rather than rational (analytical) appraisal of evidence.

In the case of the IPO-RT, reality testing is an abstract, spontaneous cognitive-perceptual process. Subsequently, individuals may lack either conscious awareness, or veridical insight into the nature of reality testing (Denovan et al., 2017b). This is especially true because metacognition encompasses two principle mechanisms, knowledge of and control of cognition (Larkin, 2009; Schneider and Artelt, 2010). Measuring cognitive processes is difficult for these reasons. This is true of metacognitive measures generally. Consequently, the relationship between subjective performance and actual performance is often weak (Rabbitt and Abson, 1990; Reid and MacLullich, 2006; Buelow et al., 2014). Hence, future studies should examine the extent to which belief in science predicts performance on objective critical thinking skills tests. This will reveal the degree to which belief in the scientific approach corresponds to an analytical thinking style.

It would also be worthwhile examining interactions between other factors related to cognitive style, such as dogmatism, and belief in science. Dogmatism is particularly pertinent because it denotes close-mindedness (Rokeach, 1960; Shearman and Levine, 2006). Specifically, the propensity to select and process information in a manner that reinforces prior opinions/expectations (Ottati et al., 2018). Accordingly, inflexible adherence to belief is likely to affect appraisal of evidence independent of thinking style. Open-minded cognition in contrast is unbiased and involves selection and processing of information in a manner unaffected by prior opinions/expectations (Church and Samuelson, 2016; Ottati et al., 2018). In the case of belief in science, this could produce overreliance on the concept of science and a dismissal of the limitations of the scientific approach. This is certainly the case when science acts as a form of faith that assists individuals to cope with stressful and anxiety-provoking situations (Farias et al., 2013). This represents an affective rather than a rational approach, which is the antithesis of analytical, objective thought. Hence, scientific extremism is a form of radical secular faith characterized by a subjective worldview.

This paper indicates that the BISS is satisfactory at a psychometric level. However, further research is necessary because belief in science is a relatively new construct. Explicitly, consideration of this alongside other belief related measures would further understanding of the belief in science construct. This is important because secular beliefs, such as Humanism and belief in progress have demonstrated the same compensatory mechanisms as belief in science (Rutjens et al., 2010). Examining relationships between these factors will provide a better understanding of their commonalities and differences. For instance, belief in science provides a framework for comprehending the world. Within this science, people may regard science as intellectually and socially progressive. However, science in the strictest sense is neutral and amoral.

Indeed, as Sarewitz (2015) notes, the social, moral, and ethical implications of deploying advances, such as new technology are contentious rather than the science findings. Thus, scientific advancements may not produce beneficial outcomes. In this context, it may prove worthwhile to investigate whether increased understanding of the scientific method reduces its positive effects relative to Humanism and belief in progress. If no differences are evident, then this suggests that any belief system that provides explanations of the world will afford comfort and assurance (see Preston and Epley, 2005). Thus, it may be that positive beliefs by their nature have beneficial psychological effects. These arise largely from subjective rather than evidential means.

ETHICS STATEMENT

The research team gained ethical authorization for a program of studies exploring relationships between anomalous beliefs, decision-making, and cognitive-perceptual personality factors as part of the grant bidding process. In total, there were three bi-annual calls (September 2012, 2014, and 2016). Review rated each application as routine and granted ethical approval. The Director of the Research Institute for Health and Social Change (Faculty of Health, Psychology and Social Care) and Ethics Committee within the Manchester Metropolitan University supervised this process. This process demanded that two experienced reviewers scrutinized the documentation. If research, as in this case, was classified as routine this constitutes full ethical approval. This was the required level of institutional approval at that point in time.

REFERENCES

Aghababaei, N., Sohrabi, F., Eskandari, H., Borjali, A., Farrokhi, N., and Chen, Z. J. (2016). Predicting subjective well-being by religious and scientific attitudes with hope, purpose in life, and death anxiety as mediators. *Pers. Indiv. Diff.* 90, 93–98. doi: 10.1016/j.paid.2015.10.046

Ano, G. G., and Vasconcelles, E. B. (2005). Religious coping and psychological adjustment to stress: a meta-analysis. *J. Clin. Psychol.* 61, 461–480. doi: 10.1002/jclp.20049

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guildford Press.

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen and J. S. Long (Beverly Hills, CA: Sage), 136–162.

Buelow, M. T., Tremont, G., Frakey, L. L., Grace, J., and Ott, B. R. (2014). Utility of the cognitive difficulties scale and association with objective test performance. *Am. J. Alzheimer's Dis. Other Dement.* 29, 755–761. doi: 10.1177/1533317514539032

Byrne, B. M. (2010). *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming*. New York, NY: Routledge/Taylor & Francis Group.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equa. Model.* 14, 464–504. doi: 10.1080/1070551070130 1834

Church, I., and Samuelson, P. (2016). *Intellectual Humility: An Introduction to the Philosophy and Science*. London: Bloomsbury Publishing.

Çokluk, Ö, and Koçak, D. (2016). Using horn's parallel analysis method in exploratory factor analysis for determining the number of factors. *Educ. Sci.* 16, 537–551. doi: 10.12738/estp.2016.2.0328

Dagnall, N., Denovan, A., Parker, A., Drinkwater, K., and Walsh, S. (2018). Confirmatory factor analysis of the inventory of personality organization-reality testing subscale. *Front. Psychol.* 9:1116. doi: 10.3389/fpsyg.2018.01116

Dagnall, N., Drinkwater, K., Denovan, A., and Parker, A. (2015a). Suggestion, belief in the paranormal, proneness to reality testing deficits and perception of an allegedly haunted building. *J. Parapsychol.* 79, 87–104.

Dagnall, N., Drinkwater, K., Parker, A., Denovan, A., and Parton, M. (2015b). Conspiracy theory and cognitive style: a worldview. *Front. Psychol.* 6:206. doi: 10.3389/fpsyg.2015.00206

Dagnall, N., Drinkwater, K., Parker, A., and Rowley, K. (2014). Misperception of chance, conjunction, belief in the paranormal and reality testing: a reappraisal. *Appl. Cogn. Psychol.* 28, 711–719. doi: 10.1002/acp.3057

Deng, L., and Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educ. Psychol. Meas.* 77, 185–203. doi: 10.1177/0013164416658325

Denovan, A., Dagnall, N., Dhingra, K., and Grogan, S. (2017a). Evaluating the perceived stress scale among UK university students: implications for stress measurement and management. *Stud. High. Educ.* 44, 120–133. doi: 10.1080/03075079.2017.1340445

Denovan, A., Dagnall, N., Drinkwater, K., Parker, A., and Clough, P. (2017b). Perception of risk and terrorism-related behavior change: dual influences of probabilistic reasoning and reality testing. *Front. Psychol.* 8:1721. doi: 10.3389/fpsyg.2017.01721

Drinkwater, K., Dagnall, N., and Parker, A. (2012). Reality testing, conspiracy theories, and paranormal beliefs. *J. Parapsychol.* 76, 57–77.

Drinkwater, K., Denovan, A., Dagnall, N., and Parker, A. (2017). An Assessment of the dimensionality and factorial structure of the revised paranormal belief scale. *Front. Psychol.* 8:1693. doi: 10.3389/fpsyg.2017.01693

AUTHOR CONTRIBUTIONS

ND contributed to theoretical focus and analysis, and design, background, and data collection. AD contributed to theoretical focus, and led on analysis and model testing. KD contributed to and supported all sections. AP commented on drafts – provided theoretical background and draft feedback.

Drinkwater, K., Denovan, A., Dagnall, N., and Parker, A. (2018). The Australian sheep-goat scale: an evaluation of factor structure and convergent validity. *Front. Psychol.* 9:1594. doi: 10.3389/fpsyg.2018.01594

Farias, M., Newheiser, A. K., Kahane, G., and de Toledo, Z. (2013). Scientific faith: belief in science increases in the face of stress and existential anxiety. *J. Exp. Soc. Psychol.* 49, 1210–1213. doi: 10.1016/j.jesp.2013.05.008

Gray, J. (2004). *Heresies*. London: Granta books.

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1998). *Multivariate Data Analysis*, 5th Edn. Upper Saddle River, NJ: Prentice Hall.

Heine, S. J., Proulx, T., and Vohs, K. D. (2006). The meaning maintenance model: On the coherence of social motivations. *Pers. Soc. Psychol. Rev.* 10, 88–110. doi: 10.1207/s15327957pspr1002_1

Hogg, M. A., and Mulling, B. (1999). "Joining groups to reduce uncertainty: Subjective uncertainty reduction and group identification," in *Social Identity and Social Cognition*, eds D. Abrams and M. A. Hogg (Oxford, UK: Blackwell), 249–279.

Hopwood, C. J., and Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Pers. Soc. Psychol. Rev.* 14, 332–346. doi: 10.1177/1088868310361240

Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., Vandenberg, R. J., et al. (1997). Exploratory and confirmatory factor analysis: guidelines, issues, and alternatives. *J. Organ. Behav.* 18, 667–683.

Inzlicht, M., Tullett, A. M., and Good, M. (2011). The need to believe: A neuroscience account of religion as a motivated process. *Relig. Brain Behav.* 1, 192–212. doi: 10.1080/2153599X.2011.647849

Irwin, H. J. (2004). Reality testing and the formation of paranormal beliefs: a constructive replication. *J. Soc. Psych. Res.* 68, 143–152.

Irwin, H. J., Dagnall, N., and Drinkwater, K. (2015). The role of doublethink and other coping processes in paranormal and related beliefs. *J. Soc. Psych. Res.* 79, 80–97.

Irwin, H. J., Dagnall, N., and Drinkwater, K. (2016). Dispositional scepticism, attitudes to science, and belief in the paranormal. *Aus. J. Parapsychol.* 16, 117–131.

Irwin, H. J., and Marks, A. D. (2013). The 'Survey of scientifically unaccepted beliefs': a new measure of paranormal and related beliefs. *Aus. J. Parapsychol.* 13, 133–167.

Kay, A. C., Whitson, J. A., Gaucher, D., and Galinsky, A. D. (2009). Compensatory control: achieving order through the mind, our institutions, and the heavens. *Curr. Dir. Psychol. Sci.* 18, 264–268. doi: 10.1111/j.1467-8721.2009.01649.x

Kernberg, O. F. (1996). "A psychoanalytic theory of personality disorders," in *Major Theories of Personality Disorder*, eds J. F. Clarkin and M. F. Lenzenweger (New York, NY: Guilford Press), 106–140.

Langdon, R., and Coltheart, M. (2000). The cognitive neuropsychology of delusions. *Mind Lang.* 15, 184–218. doi: 10.1111/1468-0017.00129

Larkin, S. (2009). *Metacognition in Young Children*. New York, NY: Routledge.

Lenzenweger, M. F., Clarkin, J. F., Kernberg, O. F., and Foelsch, P. A. (2001). The Inventory of personality organization: psychometric properties, factorial composition, and criterion relations with affect, aggressive dyscontrol, psychosis proneness, and self-domains in a nonclinical sample. *Psychol. Assess.* 13, 577–591. doi: 10.1037/1040-3590.13.4.577

Lord, C. G., Ross, L., and Lepper, M. R. (1979). Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* 37, 2098–2109. doi: 10.1037/0022-3514.37.11.2098

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* 57, 519–530. doi: 10.1093/biomet/57.3.519

Marsh, H. W., Vallerand, R. J., Lafrenière, M. A. K., Parker, P., Morin, A. J., Carbonneau, N., et al. (2013). Passion: does one scale fit all? Construct validity of two-factor passion scale and psychometric invariance over different activities and languages. *Psychol. Assess.* 25, 796–809. doi: 10.1037/a0032573

Muthén, L. K., and Muthén, B. O. (2015). *Mplus User's Guide*, 7th Edn. Los Angeles, CA: Muthén & Muthén.

Nevitt, J., and Hancock, G. R. (2001). Performance of bootstrapping approaches to model test statistics and parameter standard error estimation in structural equation modeling. *Struct. Equa. Model.* 8, 353–377.

Norman, G. R., and Streiner, D. L. (1994). *Biostatistics: The Bare Essentials*. St. Louis, MO: Mosby-Year.

Novak, T. P., and Hoffman, D. L. (2008). The fit of thinking style and situation: new measures of situation-specific experiential and rational cognition. *J. Consum. Res.* 36, 56–72. doi: 10.1086/596026

O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behav. Res. Methods Instr. compu.* 32, 396–402. doi: 10.3758/BF03200807

Ottati, V., Wilson, C., Osteen, C., and Distefano, Y. (2018). Experimental demonstrations of the earned dogmatism effect using a variety of optimal manipulations: commentary and response to Calin-Jageman (2018). *J. Exp. Soc. Psychol.* 61, 131–138. doi: 10.1016/j.jesp.2018.05.010

Pallant, J. (2007). *SPSS Survival Manual: A Step by Step Guide to Data Analysis for WINDOWS,* 3rd ed. Maidenhead, UK: Open University Press.

Paulhus, D. L., and Carey, J. M. (2010). The FAD-Plus: Measuring lay beliefs regarding free will and related constructs. *J. Pers. Assess.* 93, 96–104. doi: 10.1080/00223891.2010.528483

Preston, J., and Epley, N. (2005). Explanations versus applications: The explanatory power of valuable beliefs. *Psychol. Sci.* 16, 826–832. doi: 10.1111/j.1467-9280.2005.01621.x

Rabbitt, P., and Abson, V. (1990). 'Lost and Found': Some logical and methodological limitations of self-report questionnaires as tools to study cognitive ageing. *Br. J. Psychol.* 81, 1–16. doi: 10.1111/j.2044-8295.1990.tb02342.x

Reid, L. M., and MacLullich, A. M. (2006). Subjective memory complaints and cognitive impairment in older people. *Dement. Geriatr. Cogn. Disord.* 22, 471–485. doi: 10.1159/000096295

Rokeach, M. (1960). *The Open and Closed Mind: Investigations Into the Nature of Belief Systems and Personality Systems*. New York, NY: Basic Books.

Rutjens, B. T., van der Pligt, J., and van Harreveld, F. (2009). Things will get better: the anxiety-buffering qualities of progressive hope. *Pers. Soc. Psychol. Bull.* 35, 535–543. doi: 10.1177/0146167208331252

Rutjens, B. T., van Harreveld, F., and van der Pligt, J. (2010). Yes we can: belief in progress as compensatory control. *Soc. Psychol. Pers. Sci.* 1, 246–252. doi: 10.1177/1948550610361782

Sarewitz, D. (2015). CRISPR: science can't solve it. *Nat. News* 522:413. doi: 10.1038/522413a

Schneider, W., and Artelt, C. (2010). Metacognition and mathematics education. *ZDM* 42, 149–161. doi: 10.1007/s11858-010-0240-2

Shearman, S. M., and Levine, T. R. (2006). Dogmatism updated: a scale revision and validation. *Commun. Quar.* 54, 275–291.

Shiloh, S., Salton, E., and Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Pers. Individ. Diff.* 32, 415–429. doi: 10.1016/S0191-8869(01)00034-4

Ståhl, T., Zaal, M. P., and Skitka, L. J. (2016). Moralized rationality: Relying on logic and evidence in the formation and evaluation of belief can be seen as a moral issue. *PloS One* 11:e0166332. doi: 10.1371/journal.pone.0166332

Suhr, D. (2006). "Exploratory or Confirmatory Factor Analysis," in *Proceedings of the SAS Users Group International Conference*. Cary: SAS Institute Inc., 1–17.

Tabachnick, B. G., and Fidell, L. S. (2014). *Using Multivariate Statistics*, 6th ed. Harlow, UK: Pearson.

Valdesolo, P., Park, J., and Gottlieb, S. (2016). Awe and scientific explanation. *Emotion* 16, 937–940. doi: 10.1037/emo0000213

Van der Steen, J. T., Kruse, R. L., Szafara, K. L., Mehr, D. R., van der Wal, G., Ribbe, M. W., et al. (2008). Benefits and pitfalls of pooling datasets from comparable observational studies: combining US and Dutch nursing home studies. *Palliat. Med.* 22, 750–759. doi: 10.1177/0269216308094102

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.* 54, 1063–1070. doi: 10.1037/0022-3514.54.6.1063

Whitmarsh, L. (2011). Scepticism and uncertainty about climate change: dimensions, determinants and change over time. *Glob. Environ. Chang.* 21, 690–700. doi: 10.1016/j.gloenvcha.2011.01.016

Williams, R. N., Taylor, C. B., and Hintze, W. J. (1989). The influence of religious orientation on belief in science, religion, and the paranormal. *J. Psychol. Theol.* 17, 352–359. doi: 10.1177/009164718901700405

Ziman, J. M. (1978/1991). *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*. Cambridge: Cambridge University Press.

Zusne, L., and Jones, W. H. (1989). *Anomalistic Psychology: A Study of Magical Thinking*, 2nd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 14 May 2019
doi: 10.3389/fpsyg.2019.01125

Assessment of Entrepreneurial Orientation in Vocational Training Students: Development of a New Scale and Relationships With Self-Efficacy and Personal Initiative

Arantxa Gorostiaga[1], *Jone Aliri*[1]*, *Imanol Ulacia*[1], *Goretti Soroa*[2], *Nekane Balluerka*[1], *Aitor Aritzeta*[3] *and Alexander Muela*[2]

[1] Department of Social Psychology and Behavioral Sciences Methods, University of the Basque Country UPV/EHU, San Sebastian, Spain, [2] Department of Personality, Assessment and Psychological Treatment, University of the Basque Country UPV/EHU, San Sebastian, Spain, [3] Department of Basic Psychological Processes and Development, University of the Basque Country UPV/EHU, San Sebastian, Spain

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Javier Ortuño Sierra,
University of La Rioja, Spain
Sonja Heintz,
University of Zurich, Switzerland

****Correspondence:***
Jone Aliri
jone.aliri@ehu.eus

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *10 December 2018*
Accepted: *29 April 2019*
Published: *14 May 2019*

Citation:
Gorostiaga A, Aliri J, Ulacia I, Soroa G, Balluerka N, Aritzeta A and Muela A (2019) Assessment of Entrepreneurial Orientation in Vocational Training Students: Development of a New Scale and Relationships With Self-Efficacy and Personal Initiative.
Front. Psychol. 10:1125.
doi: 10.3389/fpsyg.2019.01125

Having emerged as an important concept in the organizational field, entrepreneurial orientation has also become a key idea in the context of education. Indeed, entrepreneurial education is now one of the common objectives for education and training systems in the European Union. Despite its importance, however, there is a scarcity of valid and reliable measures for assessing entrepreneurial orientation in students. The present study aimed to address this by developing and examining the psychometric properties of the Entrepreneurial Orientation Scale (EOS). A second objective is to study the relationships between entrepreneurial orientation and gender, self-efficacy, and personal initiative. The sample comprised 411 vocational training students (50.36% male, 49.64% female). The final version of the instrument comprised 32 items assessing six dimensions: innovativeness, risk-taking, proactiveness, competitiveness, achievement orientation, and learning orientation. The EOS showed good psychometric properties and its dimensions demonstrated concurrent relationships with self-efficacy and personal initiative. The EOS may be used to measure entrepreneurial orientation in the educational context and to evaluate interventions designed to promote an entrepreneurial spirit in schools, colleges, and universities.

Keywords: entrepreneurial orientation, self-efficacy, personal initiative, measurement invariance, multi-group confirmatory factor analysis

INTRODUCTION

Since the 1980s, increasing importance has been attached to the concept of entrepreneurial orientation (EO) (Miller, 1983; Covin and Slevin, 1989), especially in the literature on entrepreneurship and organizational performance. Various studies have sought to define this concept in terms of certain psychological, sociodemographic, and entrepreneurial profiles

(Shapero and Sokol, 1982; Lumpkin and Dess, 1996; Veciana, 1999; Krauss et al., 2005; Rauch et al., 2009; Vij and Bedi, 2012). For example, Lumpkin and Dess (1996) define EO as the processes through which organizations seek to develop a strategic basis for decisions and entrepreneurial actions. Krauss et al. (2005) emphasize the psychological nature of EO and point out that orientations, in contrast to traits, are culturally determined and influenced by context.

The first dimensions of EO to be consistently identified by organizational research were innovativeness, risk-taking, and proactiveness (Covin and Slevin, 1991). In the organizational context, innovativeness refers to the propensity toward creativity and experimentation through the introduction of new products and services, as well as to technological leadership in new processes. Risk-taking is the degree to which firms or managers are willing to consider investing in and committing resources to projects that may well fail, and to assume the risks associated with such initiatives. Finally, proactiveness is about seeking opportunities and refers to how an organization goes about anticipating future market needs. Lumpkin and Dess (1996) subsequently proposed another two dimensions of EO: competitive aggressiveness and autonomy. Competitive aggressiveness refers to the intensity of approach and head-to-head posturing that a company may need in order to compete with its rivals. The autonomy dimension reflects the independent and autonomous actions that are implemented by leaders and teams with the aim of launching a new venture. Krauss et al. (2005) later added two more elements to this framework, namely *achievement orientation* and *learning orientation*. Firms or individuals with a strong achievement orientation perform better on non-routine tasks and take responsibility for their performance. Learning orientation refers to the ability to learn from both positive and negative experiences and to the willingness to question assumptions or mental models in the pursuit of success.

Several studies have suggested that the different dimensions of EO are intercorrelated (Bhuian et al., 2005; Tan and Tan, 2005), or even that they may be subsumed under a single factor (Covin et al., 1994; Wiklund and Shepherd, 2003). However, other authors consider them to be independent aspects of a multidimensional construct (Lumpkin and Dess, 1996; George, 2011). In the meta-analysis carried out by Rauch et al. (2009), 37 of the 51 studies reviewed considered the EO construct to be unidimensional, while the remainder viewed it as multidimensional. The debate over the dimensionality of the construct therefore remains open.

Although the notion of EO emerged in the organizational context, it is now a key concept in the field of education, especially in the sphere of vocational training. This is illustrated by the fact that a "sense of initiative and entrepreneurship" is regarded by the European Commission as one of the key competences for lifelong learning (European Commission, 2007). Likewise, entrepreneurial education is one of the three key areas targeted by the Entrepreneurship 2020 Action Plan ("Promoting the spirit of entrepreneurship in schools and universities"), which the European Commission adopted in January 2013.

When the aim is to study entrepreneurial orientation in contexts other than the organizational one (e.g., the educational context), the focus needs to be on teaching and learning activities, as well as on other everyday activities. This has been done, for example, by Bolton and Lane (2011) with university students, and Kurniawan et al. (2019) with high school students.

Thus, in the present study, and drawing on existing models, we define entrepreneurial orientation as the psychological propensity of individuals to propose innovative and creative solutions to problems and to show proactiveness, autonomy, and competitiveness in the various spheres of their life, assuming the risks associated with their decisions and showing a marked orientation toward achievement and learning. Consequently, we take as our reference the seven dimensions of entrepreneurial orientation considered by Krauss et al. (2005) and apply them to a context other than the organizational one.

Research on gender differences in EO and its dimensions has yielded inconsistent results. Some authors have reported a higher level of EO among men (Bilić et al., 2011; Goktan and Gupta, 2015), although a study involving undergraduates found no such difference (Hunt, 2016). As regards the dimensions of EO, some studies have found that men score higher on innovativeness (Ayub et al., 2013; Reyes et al., 2014). However, Pérez-Quintana (2013) found no difference between men and women in this respect, and in the multi-country study by Lim and Envick (2013) a gender difference was observed in Fiji but not in the United States, Korea, or Malaysia. With regard to risk-taking, most studies have found higher scores among men (Ayub et al., 2013; Lim and Envick, 2013, in three of the four countries studied; Taatila and Down, 2012; Pérez-Quintana, 2013). However, Reyes et al. (2014) found no gender differences on the dimension which they labeled "risk propensity." For the proactiveness dimension, some studies report higher scores in women (Ayub et al., 2013; Marques et al., 2018), while others associate higher scores with men (Callaghan and Venter, 2011; Taatila and Down, 2012; Pérez-Quintana, 2013). Finally, men are generally reported to score higher on competitive aggressiveness and autonomy (Ayub et al., 2013; Lim and Envick, 2013). Given these inconsistent results regarding the relationship between gender and EO, investigating possible differences in the educational field could make a useful contribution.

Several studies have analyzed the relationship between EO and a series of variables in the literature on entrepreneurship, including self-efficacy and personal initiative. The study of these two variables is particularly relevant because there is evidence that individuals choose to become entrepreneurs most directly because they are high in self-efficacy (Zhao et al., 2005), while recent research has underlined the positive and significant association between personal initiative and social entrepreneurial behavior (Nsereko et al., 2018).

Self-efficacy is a concept that describes an individual's belief in his/her ability to succeed in a given task, and it could explain human behavior, since it plays an influential role in determining an individual's choice, level of effort, and perseverance in meeting certain objectives (Bandura, 1977; Chen et al., 2004; Sesen, 2013). In the scientific literature on entrepreneurship, researchers have tended to study the construct of entrepreneurial self-efficacy

(ESE) as a key antecedent of new venture intentions (Boyd and Vozikis, 1994). However, as McGee et al. (2009) point out, disagreement exists as to whether the ESE construct is more appropriate than general self-efficacy (GSE) for that purpose. In this respect, some studies have found that self-efficacy is positively related to EO (Hashemi et al., 2012; Arrighetti et al., 2013; Malebana and Swanepoel, 2014; Mohd et al., 2014) and that entrepreneurs score higher on self-efficacy than do non-entrepreneurs (Markman et al., 2005).

Personal initiative is defined as a set of behaviors related to proactiveness, persistence, and self-starting, which are necessary when people encounter difficulties in achieving goals (Frese and Fay, 2001). Some studies have concluded that entrepreneurs show higher levels of personal initiative than do non-entrepreneurs (Frese et al., 1997; Frese and Fay, 2001; Lisbona and Frese, 2012). Furthermore, personal initiative shows positive correlations with entrepreneurial success (Crant, 1995; Koop et al., 2000; Korunka et al., 2003; Krauss et al., 2005) and with entrepreneurial orientation (Koop et al., 2000; Krauss et al., 2005). However, these relationships have not been widely studied outside the organizational field, and more research is therefore needed.

Although instruments for assessing EO are available (Rauch et al., 2009) most of them have been developed for use in the organizational context. As regards the instruments used in the educational context, they have generally been validated with university students and have been based either on the three dimensions defined by Covin and Slevin in 1991 (e.g., Taatila and Down, 2012; Mutlutürk and Mardikyan, 2018) or on the five dimensions defined by Lumpkin and Dess, 1996 (e.g., Bolton and Lane, 2011; Vogelsang, 2015; Kurniawan et al., 2019). To date, no instrument based on the seven dimensions defined by Krauss et al. (2005) has been used in the educational field. Therefore, we consider it necessary to develop a new instrument that is based on this theoretical model and which includes the dimensions of achievement orientation and learning orientation. Furthermore, given the controversy surrounding the dimensionality of the construct, a number of authors have pointed out that the development of new instruments could make a considerable contribution to our understanding of EO (Rauch et al., 2009).

The first objective of the present study was therefore to develop a reliable and valid instrument for measuring EO, the *Entrepreneurial Orientation Scale* (EOS), and to examine its psychometric properties. More specifically, we aimed to provide evidence of its internal structure, of measurement invariance across gender groups, and of reliability of scores in terms of both internal consistency and temporal stability. Finally, we also sought to provide evidence of convergent validity.

With the aim of helping to clarify the relationships between EO and other relevant variables, the second objective was to explore latent and observed mean differences across gender and to examine the concurrent relationships of EO with self-efficacy and personal initiative. Given that the study was conducted in the educational field of vocational training, we considered that it would be more appropriate to work with the construct of GSE, rather than ESE, because vocational students do not usually have the immediate intention to start a new business.

MATERIALS AND METHODS

Participants

The sample comprised 411 students (204 female, 207 male) aged between 16 and 57 years (*M* = 22.91; *SD* = 6.26). They were recruited from across 13 vocational training colleges in the Basque Country (Spain), and were enrolled in courses at either the intermediate (17.8% of participants) or advanced (82.2% of participants) level of training. Overall, 53% of the sample had previous work experience, 34.1% had taken part in courses or activities related to entrepreneurship, and 54.3% attended publicly-funded colleges. Sampling was incidental, but in order to ensure that the sample size was sufficient for carrying out the multi-group confirmatory factor analysis (CFA) by gender, we recruited a minimum of 200 participants per group (González-Romá et al., 2006; Pendergast et al., 2017).

Instruments

Entrepreneurial Orientation Scale (EOS)

In a preliminary stage of the present study, we drew up 85 items covering the seven dimensions featured in the aforementioned theoretical model of EO. Sixty-five of these items were positively worded (i.e., stronger agreement with the statement indicated a higher level of EO), while the remainder were negatively worded. This initial battery of items was then submitted to a panel of experts who were asked to rate the relevance of the statements to the construct of EO and to indicate the dimension to which they believed each one corresponded. The panel of experts comprised four university lecturers and three enterprise project coordinators from different institutions. Based on their feedback, we selected items that fulfilled the following two criteria: mean score for relevance above 2.5 (on a scale of 1–4) and matched to the corresponding theoretical dimension by a majority of the experts. This process produced a list of 58 items.

We then piloted this preliminary measure in a sample comprising 82 vocational training students (48% male, 52% female) from three different colleges and four stages of training. Of these students, 34.1% had previous work experience. Analysis of the data obtained – both quantitative (descriptive analysis and corrected item-total correlations) and qualitative (analysis of items that students found difficult to understand) – led us to eliminate 14 items and reformulate a further five. The version of the EOS used in the present study therefore comprised 44 items, each rated on a five-point Likert-like scale (1 = *Totally agree* to 5 = *Totally disagree*). The final version of the instrument contained 32 items. Additional information about the process of developing the instrument can be found in the **Supplementary Material** (**Tables 1**, **2**).

Entrepreneurial Attitude Scale (Roth and Lacoa, 2009)

This is a unidimensional instrument consisting of 15 items (e.g., "I'm always ready to take on new projects") that are rated on a four-point Likert-like scale (1 = *Totally disagree* to 4 = *Totally agree*). The statements relate to proactiveness, propensity to excellence, effectiveness seeking, trust in success, and resilience. The instrument shows adequate psychometric properties (Roth and Lacoa, 2009). As this scale was originally

TABLE 1 | Fit indices for the CFA testing the unidimensional and six-factor models.

Models	χ^2 (*df*)	CFI	TLI	RMSEA (90% CI)
CFA 1 dim. congeneric	2412.123 (464)	0.632	0.606	0.101 (0.097–0.105)
CFA 6 dim. congeneric	875.366 (449)	0.919	0.911	0.048 (0.043–0.053)
CFA 6 dim. tau-equivalent	1172.559 (475)	0.868	0.862	0.060 (0.055–0.064)

χ^2, Chi squared; df, degrees of freedom; CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, root mean square error of approximation; CI, confidence interval.

developed for application in a Bolivian population, in a previous study small changes were made to three items so as to adapt them to the cultural context of the Basque Country (Balluerka et al., 2014). The scores obtained with this modified instrument yielded an alpha coefficient (internal consistency) of 0.92. The instrument used in the present study had a single factor and an ordinal omega coefficient (internal consistency) of 0.90 (95% CI 0.80–1.00).

Spanish Adaptation of the General Self-Efficacy Scale (Baessler and Schwarzer, 1996; Sanjuán et al., 2000)

This instrument assesses perceived personal competence in dealing effectively with a wide variety of stressful situations. It consists of 10 items (e.g., "I can solve most problems if I invest the necessary effort") that are rated on a ten-point Likert-like scale (1 = *Totally disagree* to 10 = *Totally agree*). The Spanish adaptation shows adequate psychometric properties (Sanjuán et al., 2000). The internal consistency of the score was α = 0.87 and the predictive validity indexes were good. In the present study the internal consistency was good (ordinal omega coefficient = 0.92 [95% CI 0.82–1.00]).

Scale for Measuring Personal Initiative in the Educational Field (EMIPAE, Balluerka et al., 2014)

This is a three-factor instrument consisting of 17 items. The factors are proactivity and prosocial behavior (e.g., "I usually participate actively in the classroom/workshop/laboratory, even if I do not receive anything in return"), persistence [e.g., "When I

TABLE 2 | Standardized factor loadings from the CFA of the six-factor model (*N* = 411).

Items	F1	F2	F3	F4	F5	F6
6. I like teachers with a different approach and who make use of new teaching methods.	0.75					
13. My goal is to have a job that is more about routine than creativity.	0.45					
18. I like to work and take part in groups where new or innovative ideas emerge.	0.72					
25. I like innovative teachers more than traditional ones.	0.72					
1. You have to take risks at times in order to be successful in life.		0.53				
7. I like to make risky decisions.		0.57				
8. In order to create something of value, you have to be prepared to make mistakes.		0.32				
17. I admire people who assume large risks.		0.69				
29. In order to create something of value, you need to take risks.		0.58				
5. I take the initiative whenever I have the opportunity to do so.			0.71			
16. In class I'm often the first person to propose things.			0.65			
27. I like to take the initiative in almost everything I do.			0.64			
2. I usually compete with my classmates.				0.72		
3. For me, being competitive is a good thing.				0.67		
9. Life in general is all about competition.				0.39		
19. I often strive to be better than others.				0.73		
20. I prefer not to have to compete.				0.50		
24. I like teachers who encourage competitiveness among their students.				0.69		
28. I often bet my classmates that I'm better than they are at something.				0.48		
30. I see myself becoming a businessman/woman and always competing.				0.67		
10. Before beginning a task I need to set myself some clear goals.					0.52	
11. Trying to do better (in my studies, in sport, etc.) is important to me.					0.73	
14. I get a special feeling whenever I achieve a goal (in my studies, in sport, etc.).					0.57	
23. I like to set myself goals that imply a challenge (in class, in sport, etc.).					0.73	
31. In order to achieve a goal I usually break it down into smaller objectives.					0.44	
4. My goal is to have a job where I am constantly learning new things.						0.65
12. You learn from your mistakes.						0.47
15. Life is a constant learning process.						0.69
21. I like people who never stop learning.						0.76
22. I try to learn new things every day.						0.72
26. For a company to be successful, its employees have to be learning all the time.						0.55
32. I always try to learn from my experiences.						0.69

Original items were in Spanish, their English translation is provided.

no longer understand the contents of a module/project/subject, I get frustrated and give up" (reverse-scored item)], and self-starting (e.g., "I am particularly good at putting into practice the ideas I had in the classroom/workshop/laboratory"). The items are rated on a five-point Likert-like scale (1 = *Totally disagree* to 5 = *Totally agree*). The instrument shows adequate psychometric properties (Balluerka et al., 2014). Internal consistency indexes ($\alpha_{\text{proactivity}}$ = 0.72, $\alpha_{\text{persistence}}$ = 0.73, and $\alpha_{\text{self-starting}}$ = 0.57) were acceptable and the scores showed evidence of convergent validity and criterion validity. Scores in the present study yielded satisfactory internal consistency indices ($\text{omega}_{\text{proactivity}}$ = 0.87 [95% CI 0.76–0.96], $\text{omega}_{\text{persistence}}$ = 0.86 [95% CI 0.78–0.94], and $\text{omega}_{\text{self-starting}}$ = 0.74 [95% CI 0.63–0.85]).

Sociodemographic Data Sheet

This was developed *ad hoc* for the present study in order to collect data on gender, age, the college where students were enrolled, level of studies (intermediate or advanced), course year, previous work experience, and profession (in the case of previous experience).

Procedure

The 44-item version of the EOS and the instruments required for its validation were administered to participants. The order of administration was as follows: Sociodemographic data sheet, the EOS, the EMIPAE, the Entrepreneurial Attitude Scale, and the GSE Scale. The study was approved by the Research and Teaching Ethics Committee of the University of the Basque Country. In accordance with the Declaration of Helsinki, written informed consent was sought from the heads of the training colleges, from the parents or legal guardians of students who were still minors, and from participants themselves.

Data Analysis

In order to select the items that would be included in the validated version of the EOS we calculated corrected item-total correlations within each dimension. Items were retained if they achieved a corrected item-total correlation of 0.30 or higher. The criterion for maintaining a dimension was that at least three items yielded a correlation of at least 0.30.

The selected items were then subjected to different models of CFA. The estimator used was weighted least squares mean and variance adjusted (WLSMV), and the fit indices employed were the comparative fit index (CFI) the Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA). In the case of the CFI and the TLI, values above 0.90 indicate acceptable fit. For the RMSEA, values below 0.08 indicate acceptable fit and those below 0.06 a good fit (Hu and Bentler, 1999). Factor invariance across gender groups was assessed by means of multi-group confirmatory factor analysis (MG-CFA). The fit indices of the two nested models (the configural invariance model and the scalar invariance model) were compared using the DIFFTEST procedure in order to check that they were not significantly worse in the more restrictive model.

In order to assess the reliability of EOS scores in terms of internal consistency we calculated the ordinal omega coefficient (Gadermann et al., 2012) for each dimension of the instrument; this measure was used as the tau-equivalence required by the alpha coefficient could not be assumed. The temporal stability of EOS scores was evaluated by means of the Spearman rho correlation coefficient. It should be noted that temporal stability was examined in a sub-sample of 65 participants using a 2 weeks interval between test administrations.

In order to obtain evidence of convergent validity we calculated Spearman rho correlation coefficients between the scores obtained by participants on the various dimensions of the EOS and their scores on the Entrepreneurial Attitude Scale (Roth and Lacoa, 2009).

Next, we examined whether there were gender differences in the latent and observed means for each of the dimensions. For the comparison of latent means we constrained the latent mean of the "males" group to 0. Statistical significance was determined on the basis of the z-statistic, and the effect size was estimated according to the guidelines proposed by Hancock (2001). In order to test whether the differences in latent means were also found in the observed means we computed observed mean differences (t-statistic) and their corresponding effect size (Cohen's d).

Finally, hierarchical multiple regression analyses were performed with the aim of testing the concurrent relationships of EO with GSE and the three dimensions of personal initiative. In these analyses the demographic variables gender, age, and previous work experience were controlled, and thus they were entered in the first step of the regression. In the second step, the demographic variables and all the EO dimensions were entered in the models. In each step, adjusted R squared was calculated. In the second step we also calculated the change in adjusted R squared as a measure of the effect size of the concurrent relationship between EO dimensions and self-efficacy and personal initiative. In addition, zero-order correlations among all variables used in the study were computed. The results can be seen in **Supplementary Material** (**Table 3**).

The analyses were performed using SPSS v23 and Mplus v7.4. Missing data (less than 5%) were handled using the single mean imputation procedure.

RESULTS

Dimensional Structure

Based on the corrected item-total correlations for the items in each dimension the definitive scale comprised 32 items pertaining to six of the seven dimensions originally proposed: innovativeness, 4 items (e.g., "I like to work and take part in groups where new or innovative ideas emerge"); risk-taking, 5 items (e.g., "In order to create something of value, you need to take risks"); proactiveness, 3 items (e.g., "In class I'm often the first person to propose things"); competitiveness, 8 items (e.g., "I usually compete with my classmates"); achievement orientation, 5 items (e.g., "Before beginning a task I need to set myself some clear goals"); and learning orientation, 7 items (e.g., "My goal is to have a job where I am constantly learning new things"). The autonomy dimension was eliminated as only one of its items had a corrected item-total correlation above the established cut-off.

TABLE 3 | Fit indices of the models tested to assess measurement invariance across gender groups.

Invariance model	Model constraints	χ^2 (*df*)	CFI	TLI	RMSEA (90% CI)
M1: Configural invariance	Equivalent model	1353.43 (898)***	0.911	0.902	0.050 (0.044–0.055)
M2: Metric invariance	Equivalent unstandardized factor loadings	1356.51 (924)***	0.916	0.910	0.048 (0.042–0.053)
M3: Scalar invariance	Equivalent thresholds	1444.66 (1006)***	0.915	0.916	0.046 (0.041–0.051)
Differences in model fit (M1–M3)	–	129.01 (108)	−0.004	–	0.004

χ^2, Chi squared; df, degrees of freedom; CFI, comparative fit index; TLI, Tucker-Lewis index; RMSEA, root mean square error of approximation; CI, confidence interval.
****$p < 0.001$.*

The unidimensional CFA did not show an adequate fit (see **Table 1**). However, as can be seen in **Table 1** the fit of the six-factor structure was adequate. We also tested a third model in order to determine whether tau-equivalence could be assumed. This model did not show an adequate fit. The factor loadings corresponding to the second (six-factor) model are shown in **Table 2**. Loadings for all but two of the items were both statistically significant and above 0.40. Observed and latent correlations among the six dimensions can be found in the **Supplementary Material** (**Table 4**).

Table 3 shows the results from the analysis of factor invariance of the EOS across gender groups. The constrained model with equivalent thresholds and factor loadings for males and females (scalar invariance) showed an adequate fit (CFI = 0.915; TLI = 0.916; RMSEA = 0.046), and ΔCFI ≤ 0.01 (0.911–0.915 = −0.004).

Reliability and Convergent Validity

The ordinal omega coefficients and their confidence intervals are shown in **Table 4**. These coefficients ranged between 0.68 and 0.84. The test-retest correlation coefficients (Spearman rho) ranged between 0.60 and 0.69 (see **Table 4**).

The correlation coefficients (Spearman rho) between the participants' scores on the six dimensions of the EOS and their scores on the Entrepreneurial Attitude Scale were as follows: innovativeness, 0.41; risk-taking, 0.37; proactiveness, 0.56; competitiveness, 0.34; achievement orientation, 0.54; and learning orientation, 0.55 ($p = 0.001$).

Differences in Entrepreneurial Orientation Across Gender Groups

Having established the scalar invariance of the EOS across gender groups we then compared the means – both latent and observed – obtained by males and females on the six dimensions of the scale. It can be seen in **Table 5** that although there were significant differences between males and females on the competitiveness and learning orientation dimensions, the effect sizes for all the comparisons were small.

Concurrent Relationships of EO With Self-Efficacy and Personal Initiative

Gender, age, and previous work experience accounted for 1.5% of the variance in self-efficacy. The dimensions of EO accounted for a further 26.5% (large effect size), leading to a total explained variance of 28% (see **Table 6**). Proactiveness, competitiveness, and learning orientation were significant predictors of self-efficacy. Higher scores on these EO dimensions were related to greater self-efficacy.

With respect to proactive and prosocial behavior (i.e., the first dimension of personal initiative), gender, age, and work experience explained 7.7% of its variance. An additional 25.7% was explained by the EO dimensions (large effect size), leading to a total explained variance of 33.4% (see **Table 6**). The only significant demographic predictor was gender, with females scoring higher on proactive and prosocial behavior. All the dimensions of EO, except competitiveness, were significant predictors of this outcome. Specifically, and as indicated by the beta values, higher scores on innovativeness, proactiveness, achievement orientation, and learning orientation were associated with greater proactive and prosocial behavior. Conversely, higher scores on risk-taking were related to lower scores on proactive and prosocial behavior.

The demographic variables explained 1.3% of the variance in persistence. An additional 13.7% was explained by the EO dimensions (medium effect size), leading to a total explained variance of 15% (see **Table 6**). In addition to age (demographic variable), the EO dimensions of innovativeness, risk-taking, proactiveness, and learning orientation were significant predictors of persistence. Specifically, participants scored higher on persistence with increasing age, innovativeness, proactiveness, and learning orientation. With respect to risk-taking, persistence decreased as scores on this dimension increased.

Finally, gender, age, and work experience explained 2.3% of the variance in self-starting. An additional 38.2% was explained by the EO dimensions (large effect size), leading to a total explained variance of 40.5% (see **Table 6**). All the dimensions of EO, except innovativeness, were significant predictors of self-starting. The beta values indicate that higher scores on

TABLE 4 | Reliability indices of the EOS.

Dimension	Mean (*SD*)	Omega (95% CI)	Test-retest correlation
Innovativeness	4.01 (0.63)	0.75 (0.63–0.86)	0.68***
Risk-taking	3.94 (0.51)	0.68 (0.56–0.80)	0.60***
Proactiveness	3.37 (0.69)	0.71 (0.59–0.82)	0.64***
Competitiveness	2.82 (0.71)	0.83 (0.73–0.93)	0.69***
Achievement orientation	3.89 (0.58)	0.74 (0.60–0.87)	0.63***
Learning orientation	4.45 (0.42)	0.84 (0.73–0.93)	0.61***

****$p < 0.001$.*

TABLE 5 | Differences between males and females in latent and observed means.

	Latent mean analyses		Observed mean analyses					
			Females		Males			
	z	*d*	*M*	*SD*	*M*	*SD*	*t*	*d*
Innovativeness	0.38	0.03	4.02	0.63	4.00	0.62	0.28	0.03
Risk-taking	0.44	0.02	3.95	0.50	3.93	0.53	0.47	0.05
Proactiveness	−0.41	0.02	3.35	0.70	3.38	0.68	−0.42	0.04
Competitiveness	−3.62**	0.22	2.68	0.66	2.96	0.73	−4.06**	0.40
Achievement orientation	−0.32	0.01	3.90	0.57	3.88	0.59	0.33	0.03
Learning orientation	2.93**	0.20	4.52	0.39	4.39	0.44	3.25**	0.32

**$p < 0.01$.

TABLE 6 | Multiple regressions of control variables and EO dimensions on self-efficacy and personal initiative dimensions.

	Self-efficacy[a]		Proactive and prosocial behavior[b]		Persistence[c]		Self-starting[d]	
	β	*t*	β	*t*	β	*t*	β	*t*
STEP 1								
Gender (0 = Female; 1 = Male)	0.02	0.44	−0.21	−4.36**	−0.03	−0.63	−0.02	−0.42
Age	−0.01	−0.23	0.09	1.68	0.13	2.40*	0.08	1.49
Work experience (0 = No experience; 1 = Experience)	0.15	2.83**	0.14	2.72**	0.02	0.33	0.12	2.29*
STEP 2								
Gender (0 = Female; 1 = Male)	0.03	0.68	−0.18	−4.25**	−0.01	−0.30	−0.02	−0.61
Age	−0.03	−0.63	0.04	0.86	0.10	1.99*	0.06	1.34
Work experience (0 = No experience; 1 = Experience)	0.04	0.79	0.05	1.12	−0.05	−0.91	−0.01	−0.25
Innovativeness	0.08	1.59	0.16	3.42**	0.12	2.31*	0.06	1.25
Risk-taking	−0.02	−0.36	−0.13	−2.76**	−0.17	−3.11**	−0.11	−2.54*
Proactiveness	0.27	5.18**	0.16	3.12**	0.12	2.12*	0.32	6.77**
Competitiveness	0.11	2.30*	0.02	0.49	0.08	1.52	0.16	3.64**
Achievement orientation	0.10	1.89	0.24	4.59**	0.08	1.36	0.22	4.43**
Learning orientation	0.21	3.73**	0.21	3.93**	0.23	3.89**	0.21	4.07**

[a] $R_{adj}^{2} = 0.015$ for Step 1 ($p = 0.027$), $R_{adj}^{2} = 0.280$ for Step 2 ($p < 0.001$), $\Delta R^{2} = 0.265$. [b] $R_{adj}^{2} = 0.077$ for Step 1 ($p < 0.001$), $R_{adj}^{2} = 0.334$ for Step 2 ($p < 0.001$), $\Delta R^{2} = 0.257$. [c] $R_{adj}^{2} = 0.013$ for Step 1 ($p = 0.040$), $R_{adj}^{2} = 0.150$ for Step 2 ($p < 0.001$), $\Delta R^{2} = 0.137$. [d] $R_{adj}^{2} = 0.023$ for Step 1 ($p = 0.006$), $R_{adj}^{2} = 0.405$ for Step 2 ($p < 0.001$), $\Delta R^{2} = 0.382$. *$p < 0.05$; **$p < 0.01$.

proactiveness, competitiveness, achievement orientation, and learning orientation were associated with a higher self-starting score. Again, an increase in risk-taking was related to a lower score on this dimension of personal initiative.

DISCUSSION

The first aim of this study was to develop an instrument for assessing entrepreneurial orientation and to examine its psychometric properties in the educational context. The resulting Entrepreneurial Orientation Scale (EOS) comprised 32 items distributed across six dimensions (one of the seven dimensions originally considered, namely autonomy, was eliminated). Given the debate regarding the construct of entrepreneurial orientation we tested both a unidimensional model and a multidimensional (six-factor) model and found that the latter showed the best fit. As to why the autonomy dimension did not function adequately in the educational context, a possible explanation is that, in contrast to the organizational context in which entrepreneurial orientation has traditionally been assessed, autonomy is not an aspect that is widely addressed in the context of our country's education system. It is worth remembering that in the organizational context, autonomy refers to the independent actions that are implemented by leaders and teams with the aim of launching a new venture (Lumpkin and Dess, 1996). A similar result to ours was obtained in the study by Bolton and Lane (2011), who found that the items designed to measure autonomy did not load on an independent factor, leading them to conclude that autonomy may be a characteristic that, among students, has yet to become consolidated. In a similar vein, Kurniawan et al. (2019) pointed out that the autonomy dimension is not correlated with entrepreneurial intention and therefore it lacks external validity. It should also be noted that other instruments (see, for example, Sánchez, 2010; Bolton and Lane, 2011; Taatila and Down, 2012; Ismail et al., 2015) do not include the achievement orientation and learning orientation dimensions that form part of the EOS, both of which are particularly relevant to the educational setting. Consequently, we believe that the EOS can provide a more comprehensive assessment of entrepreneurial orientation in the academic context.

Importantly, scores on the EOS showed measurement invariance across gender groups, which is a prerequisite for an analysis of differences in mean scores obtained by males and females. The scores also showed adequate reliability in terms of both temporal stability and internal consistency. In addition, the correlations with respect to the Entrepreneurial Attitude Scale may be considered as evidence of good convergent validity. The highest correlation coefficients were those for proactiveness, achievement orientation, and learning orientation, which is what one would expect given that the items of the Entrepreneurial Attitude Scale refer to proactiveness, propensity to excellence, effectiveness seeking, trust in success, and resilience.

The second objective of this study was to explore latent and observed mean differences across gender and to examine the concurrent relationships of EO with self-efficacy and personal initiative. Although gender differences in entrepreneurial orientation have been examined with other instruments, the EOS is the first for which the equivalence of the factor structures, the factor loadings, and the thresholds have been analyzed for males and females. In our study, conducted in the educational context, we found no significant differences between male and female students on four of the six dimensions, and the effect sizes for all the comparisons were small. These results are consistent with those reported by Hunt (2016) for the general construct of entrepreneurial orientation in a sample of undergraduates, as well as with the findings of Pérez-Quintana (2013) and Lim and Envick (2013) with respect to the innovativeness dimension, and with those of Reyes et al. (2014) in relation to risk-taking, once again with samples of university students. These results suggest that the gender differences observed in the organizational context are not present in the same way among students. It should also be noted that, as would be expected due to scalar invariance, we obtained practically the same results when analyzing gender differences using latent and observed scores. This suggests that the EOS has low measurement error and, therefore, that applied researchers may work with observed variables when using the instrument.

Our study, conducted in the educational field, revealed a relationship between EO and self-efficacy, which is consistent with the results obtained by Mohd et al. (2014) in the organizational setting, and by Sesen (2013) with university students. Specifically, we found that the EO dimensions of proactiveness, competitiveness, and learning orientation explained a considerable part of the variance in self-efficacy.

Regarding personal initiative, which is considered one of the eight key competencies for personal development, active citizenship, social inclusion, and employment (European Commission, 2007), EO dimensions showed large concurrent relationships, especially in relation to self-starting. The EO dimensions that predicted all three dimensions of personal initiative were proactiveness, learning orientation, and risk-taking. The negative sign of the relationship between risk-taking and personal initiative was initially surprising, since it indicated that after controlling for demographic variables and the other EO dimensions, a stronger risk-taking orientation was related to less personal initiative. However, an in-depth analysis of the characteristics of the assessment instruments used revealed that the items comprising the risk-taking dimension do not, unlike those for the other dimensions, make reference to the classroom or the educational field, but rather refer more broadly to various aspects of life (see, in **Table 2**, the content of items 1, 7, 8, 17, and 29). This is important because the instrument used to assess personal initiative refers clearly to the classroom context. At all events, the standardized coefficient of this variable in the explanatory model is the smallest in two of the three dimensions of personal initiative. Finally, it should be noted that the relationship between proactiveness and personal initiative is congruent with studies conducted in organizational settings (Koop et al., 2000; Krauss et al., 2005).

One of the limitations of the present study concerns the sole use of self-report measures, such that the results may be affected by single-method bias. In addition, all the participants came from the same geographical region. Future studies should aim to use other types of measures and to recruit more heterogeneous samples. Another limitation is that we did not test the incremental validity of the EOS in comparison with other published EO measures. This would be an important step in future research with the EOS.

Despite these limitations, we believe that the development and validation of an instrument for assessing, in the educational context, six dimensions of the construct of entrepreneurial orientation makes an important contribution to the field. The results support the multidimensional nature of this construct, which to date has not been examined with vocational training students who will shortly be entering the labor market. A further strength of our study is that we examined measurement invariance across gender groups. The instrument presented here may be used to evaluate initiatives designed to promote an entrepreneurial spirit in schools, colleges, and universities and it therefore provides added value to future research and applications.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations and the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments. The protocol was approved by the Research and Teaching Ethics Committee of the University of the Basque Country. Informed consent was sought from the heads of the training colleges, from the parents or legal guardians of students who were still minors and from participants themselves in accordance with the Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

AG, IU, and AM analyzed the theoretical framework of entrepreneurial orientation, designed the study, and wrote the first draft of the manuscript. JA and NB analyzed the data and wrote the Materials and Methods and Results section. GS and AA collected the data. All authors contributed to manuscript revision and proofreading and approved the submitted version of the manuscript.

FUNDING

This work was supported by the Basque Government (Grant No. IT892-16) and the Provincial Council of Gipuzkoa (Department of Innovation, Rural Development, and Innovation, co-funded by the Social Fund (50%) and the European Regional Development Fund (50%); Grant No. OF-208/2014-B).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01125/full#supplementary-material

REFERENCES

Arrighetti, A., Caricati, L., Landini, F., and Monacelli, N. (2013). *Explaining Entrepreneurial Orientation Among University Students: Evidence from Italy*. C.MET Working paper 01/2013. Italy: University of Parma.

Ayub, A., Razzaq, A., Aslam, M. S., and Iftakhar, H. (2013). Gender effects on entrepreneurial orientation and value innovation: evidence from Pakistan. *Eur. J. Bus. Soc. Sci.* 2, 82–90.

Baessler, J., and Schwarzer, R. (1996). Evaluación de la autoeficacia: adaptación española de la escala de autoeficacia general [assessing self-efficacy: spanish adaptation of the general self-efficacy scale]. *Ansiedad y Estrés* 2, 1–8.

Balluerka, N., Gorostiaga, A., and Ulacia, I. (2014). Assessing personal initiative among vocational training students: development and validation of a new measure. *Span. J. Psychol.* 17, 1–9. doi: 10.1017/sjp.2014.80

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychol. Rev.* 84, 191–215. doi: 10.1037/0033-295X.84.2.191

Bhuian, S. N., Menguc, B., and Bell, S. J. (2005). Just entrepreneurial enough: the moderating effect of entrepreneurship on the relationship between market orientation and performance. *J. Bus. Res.* 58, 9–17. doi: 10.1016/S0148-2963(03)00074-2

Bilić, I., Prka, A., and Vidović, G. (2011). How does education influence entrepreneurship orientation? Case study of Croatia. *Manag. J. Contemp. Manag.* 16, 115–128.

Bolton, D. L., and Lane, M. D. (2011). Individual entrepreneurial orientation: development of a measurement instrument. *Educ. Train* 54, 219–233. doi: 10.1108/00400911211210314

Boyd, N. G., and Vozikis, G. S. (1994). The influence of self-efficacy on the development of entrepreneurial intentions and actions. *Entrep. Theor. Pract.* 18, 63–77. doi: 10.1177/104225879401800404

Callaghan, C., and Venter, R. (2011). An investigation of the entrepreneurial orientation, context and entrepreneurial performance of inner-city Johannesburg street traders. *South. Afr. Bus. Rev.* 15, 28–48.

Chen, G., Gully, M. S., and Eden, D. (2004). General self-efficacy and self-esteem: toward theoretical and empirical distinction between correlated self-evaluations. *J. Organ. Behav.* 25, 375–395. doi: 10.1002/job.251

Covin, J. G., and Slevin, D. P. (1989). Strategic management of small firms in hostile and benign environments. *Strategic Manage. J.* 10, 75–87. doi: 10.1002/smj.4250100107

Covin, J. G., and Slevin, D. P. (1991). A conceptual model of entrepreneurship as firm behavior. *Entrep. Theor. Pract.* 16, 7–25. doi: 10.1177/104225879101600102

Covin, J. G., Slevin, D. P., and Schultz, R. L. (1994). Implementing strategic missions: effective strategic, structural, and tactical choices. *J. Manage. Stud.* 31, 481–503. doi: 10.1111/j.1467-6486.1994.tb00627.x

Crant, J. M. (1995). The proactive personality scale and objective job performance among real estate agents. *J. Appl. Psychol.* 80, 532–537. doi: 10.1037/0021-9010.80.4.532

European Commission (2007). *Key Competences for Lifelong Learning - European Reference Framework*. Luxembourg: Office for Official Publications of the European Communities.

Frese, M., and Fay, D. (2001). "Personal initiative (PI): An active performance concept for work in the 21st century," in *Research in Organizational Behavior*, eds B. M. Staw and R. M. Sutton (Amsterdam: Elsevier Science), 133–187. doi: 10.1016/S0191-3085(01)23005-6

Frese, M., Fay, D., Hilburger, T., Leng, K., and Tag, A. (1997). The concept of personal initiative: operationalization, reliability and validity in two german samples. *J. Occup. Organ. Psych.* 70, 139–161. doi: 10.1111/j.2044-8325.1997.tb00639.x

Gadermann, A. M., Guhn, M., and Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Pract. Assess. Res. Eval.* 17, 1–13.

George, B. A. (2011). Entrepreneurial orientation: a theoretical and empirical examination of the consequences of differing construct representations. *J. Manag. Stud.* 48, 1291–1313. doi: 10.1111/j.1467-6486.2010.01004.x

Goktan, A. B., and Gupta, V. K. (2015). Sex, gender, and individual entrepreneurial orientation: evidence from four countries. *Int. Entrep. Manag. J.* 11, 95–112. doi: 10.1007/s11365-013-0278-z

González-Romá, V., Hernández, A., and Gómez-Benito, J. (2006). Power and type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivar. Behav. Res.* 41, 29–53. doi: 10.1207/s15327906mbr4101_3

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika* 66, 373–388. doi: 10.1007/BF02294440

Hashemi, A. M. K., Hosseini, S. M., and Rezvanfar, A. (2012). Explaining entrepreneurial intention among agricultural students: effects of entrepreneurial self-efficacy and college entrepreneurial orientation. *Res. J. Bus. Manage.* 6, 91–102.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Modeling* 6, 1–55. doi: 10.1080/10705519909540118

Hunt, K. A. (2016). Individual Entrepreneurial Orientation, Gender, and Involvement in Entrepreneurial Activities: A Study of Faculty at Pennsylvania's Public Universities. Order No. 10036405). Available from ProQuest Dissertations & Theses Global. (1789310144). Available at: https://search.proquest.com/docview/1789310144?accountid=17248 (accessed March 3, 2017).

Ismail, K., Anuar, M. A., Wan Omar, W. Z., Aziz, A. A., Seohod, K., and Akhtar, C. S. (2015). Entrepreneurial intention, entrepreneurial orientation of faculty and students towards commercialization. *Procedia Soc. Behav. Sci.* 181, 349–355. doi: 10.1016/j.sbspro.2015.04.897

Koop, S., De Reu, T., and Frese, M. (2000). "Sociodemographic factors, entrepreneurial orientation, personal initiative and environmental problems in Uganda," in *Success and Failure of Microbusiness Owners in Africa: A Psychological Approach*, ed. M. Frese (Westport, CT: Quorum), 55–76.

Korunka, C., Frank, H., Lueger, M., and Mugler, J. (2003). The entrepreneurial personality in the context of resources, environment, and the start-up process: a configurational approach. *Entrep. Theor. Pract.* 28, 23–42. doi: 10.1111/1540-8520.00030

Krauss, S., Frese, M., Friedrich, C., and Unger, J. M. (2005). Entrepreneurial orientation: a psychological model of success among southern african small business owners. *Eur. J. Work Organ. Psychol.* 14, 315–344. doi: 10.1080/13594320500170227

Kurniawan, J. E., Setiawan, J. L., Sanjaya, E. L., Wardhani, F. P. I., Virlia, S., Dewi, K., et al. (2019). Developing a measurement instrument for high school students' entrepreneurial orientation. *Cogent Ed.* 6:1564423. doi: 10.1080/2331186X.2018.1564423

Lim, S., and Envick, B. R. (2013). Gender and entrepreneurial orientation: a multi-country study. *Int. Entrep. Manag. J.* 9, 465–482. doi: 10.1007/s11365-011-0183-2

Lisbona, A., and Frese, M. (2012). *Iniciativa Personal. Cómo Hacer que las Cosas Sucedan [Personal Initiative. How to make things Happen]*. Madrid: Ediciones Pirámide.

Lumpkin, G. T., and Dess, G. G. (1996). Clarifying the entrepreneurial orientation construct and linking it to performance. *Acad. Manage. Rev.* 21, 135–172. doi: 10.2307/258632

Malebana, M. J., and Swanepoel, E. (2014). The relationship between exposure to entrepreneurship education and entrepreneurial self-efficacy. *South. Afr. Bus. Rev.* 18, 1–26.

Markman, G. D., Baron, R. A., and Balkon, D. B. (2005). Are perseverance and self-efficacy costless? Assessing entrepreneurs' regretful thinking. *J. Organ. Behav.* 26, 1–19. doi: 10.1002/job.305

Marques, C., Santos, G., Galvão, A., Mascarenhas, C., and Justino, E. (2018). Entrepreneurship education, gender and family background as antecedents on the entrepreneurial orientation of university students. *Int. J. Innov. Sci.* 10, 58–70. doi: 10.1108/IJIS-07-2017-0067

McGee, J. E., Peterson, M., Mueller, S. L., and Sequeira, J. M. (2009). Entrepreneurial self-efficacy: refining the measure. *Entrep. Theor. Pract.* 33, 965–988. doi: 10.1111/j.1540-6520.2009.00304.x

Miller, D. (1983). The correlates of entrepreneurship in three types of firms. *Manage. Sci.* 29, 770–791. doi: 10.1287/mnsc.29.7.770

Mohd, R., Kamaruddin, B. H., Hassan, S., Muda, M., and Yahya, K. K. (2014). The important role of self-efficacy in determining entrepreneurial orientations of Malay small scale entrepreneurs in Malaysia. *Int. J. Manage. Stud.* 21, 61–82.

Mutlutürk, M., and Mardikyan, S. (2018). Analysing factors affecting the individual entrepreneurial orientation of university students. *J. Entrep. Educ.* 21, 1–15.

Nsereko, I., Balunywa, W., Munene, J., Orobia, L., and Muhammed, N. (2018). Personal initiative: its power in social entrepreneurial venture creation. *Cogent Bus. Manage.* 5, 1–15. doi: 10.1080/23311975.2018.1443686

Pendergast, L. E., von der Embse, N., Kilgus, S. P., and Eklund, K. R. (2017). Measurement equivalence: a non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *J. Sch. Psychol.* 60, 65–82. doi: 10.1016/j.jsp.2016.11.002

Pérez-Quintana, A. (2013). *La influencia de los Estereotipos de género en el emprendimiento: Una aplicación en el contexto de Catalunya [The Influence of Gender Stereotypes on Entrepreneurship: An Application in the Context of Catalonia]*. Ph.D. thesis, University of Barcelona, Barcelona.

Rauch, A., Wiklund, J., Lumpkin, G. T., and Frese, M. (2009). Entrepreneurial orientation and business performance: an assessment of past research and suggestions for the future. *Entrep. Theor. Pract.* 33, 761–787. doi: 10.1111/j.1540-6520.2009.00308.x

Reyes, L. E., Pinillos, M. J., and Soriano, I. (2014). Gender differences in entrepreneurial orientation. *Esic Market Econ. Bus. J.* 45, 421–439. doi: 10.1007/s10964-013-9930-8

Roth, E., and Lacoa, D. (2009). Análisis psicológico del emprendimiento en estudiantes universitarios: medición, relaciones y predicción [Psychological analysis of entrepreneurial spirit among university students: measurement, relationships, and prediction]. *Rev. Electrón. Psicol.* 7, 1–38.

Sánchez, J. C. (2010). Evaluación de la personalidad emprendedora: validez factorial del cuestionario de orientación emprendedora (COE) [Assessing entrepreneurial personality: factor validity of the entrepreneurial orientation questionnaire (COE)]. *Rev. Lat. Am. Psicol.* 42, 41–52.

Sanjuán, P., Pérez, A., and Bermúdez, J. (2000). Escala de autoeficacia general: datos psicométricos de la adaptación para la población española [General self-efficacy scale: psychometric data for the spanish adaptation]. *Psicothema* 12, 509–513.

Sesen, H. (2013). Personality or environment? A comprehensive study on the entrepreneurial intentions of university students. *Educ. Train.* 55, 624–640. doi: 10.1108/ET-05-2012-0059

Shapero, A., and Sokol, L. (1982). "The Social Dimensions of Entrepreneurship," in *The Encyclopedia of Entrepreneurship*, eds C. Kent, D. Sexton, and K. H. Vesper (Englewood Cliffs, NJ: Prentice-Hall), 72–90.

Taatila, V., and Down, S. (2012). Measuring entrepreneurial orientation of university students. *Educ. Train.* 54, 744–760. doi: 10.1108/00400911211274864

Tan, J., and Tan, D. (2005). Environment-strategy coevolution and coalignment: a staged-model of Chinese SOEs under transition. *Strategic Manage. J.* 26, 141–157. doi: 10.1002/smj.437

Veciana, J. M. (1999). Creación de empresas como programa de investigación científica [Creating companies as a scientific research program]. *Rev. Eur. Dir. Econ. Empresa* 8, 11–36.

Vij, S., and Bedi, H. S. (2012). Relationship between entrepreneurial orientation and business performance: a review of literature. *IUP J. Bus. Strategy* 9, 17–31.

Vogelsang, L. (2015). *Individual Entrepreneurial Orientation: an Assessment of Students*. Master thesis, Humboldt State University, California.

Wiklund, J., and Shepherd, D. (2003). Knowledge-based resources, entrepreneurial orientation, and the performance of small and medium sized businesses. *Strategic Manage. J.* 24, 1307–1314. doi: 10.1002/smj.360

Zhao, H., Seibert, S. E., and Hills, G. E. (2005). The mediating role of self-efficacy in the development of entrepreneurial intentions. *J. Appl. Psychol.* 90, 1265–1272. doi: 10.1037/0021-9010.90.6.1265

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 31 May 2019
doi: 10.3389/fpsyg.2019.01237

Factorial Invariance of the 10-Item Connor-Davidson Resilience Scale Across Gender Among Chinese Elders

Meng Meng[1,2†], *Jiayue He*[3†], *Yuzhu Guan*[1,2], *Haofei Zhao*[3], *Jinyao Yi*[3], *Shuqiao Yao*[3*] and *Lezhi Li*[1,2*]

[1]*Department of Nursing, Second Xiangya Hospital, Central South University, Changsha, China,* [2]*Xiangya School of Nursing, Central South University, Changsha, China,* [3]*Medical Psychological Center, Second Xiangya Hospital, Central South University, Changsha, China*

OPEN ACCESS

Edited by:
Elisa Pedroli,
Istituto Auxologico Italiano (IRCCS),
Italy

Reviewed by:
José Manuel García-Fernádez,
University of Alicante, Spain
João P. Marôco,
Higher Institute of Applied
Psychology (ISPA), Portugal

Correspondence:
Shuqiao Yao
shuqiaoyao@csu.edu.cn
Lezhi Li
lilezhi@csu.edu.cn

†These authors have contributed equally to this work and are co-first authors

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal
Frontiers in Psychology

Received: *18 February 2019*
Accepted: *10 May 2019*
Published: *31 May 2019*

Citation:
Meng M, He J, Guan Y, Zhao H, Yi J, Yao S and Li L (2019) Factorial Invariance of the 10-Item Connor-Davidson Resilience Scale Across Gender Among Chinese Elders. Front. Psychol. 10:1237. doi: 10.3389/fpsyg.2019.01237

Resilience plays an important role in the health of the elderly. The 10-item Connor-Davidson Resilience Scale (CD-RISC-10) is widely used to evaluate resilience, but its factorial invariance has not been evaluated in the Chinese elders. In the current study, 1,238 Chinese elders aged 60 years and above completed the Chinese CD-RISC-10, yielding good reliability (Cronbach's $\alpha = 0.936$, Omega coefficient = 0.83, and test-retest reliability coefficient of 0.665 after 6 months). Confirmatory factor analysis indicated that a single-factor model fitted our CD-RISC-10 data well, both for the total sample and for each gender group. Furthermore, factorial invariance across genders was supported by multigroup confirmatory factor analysis. Finally, the current study revealed greater resilience levels in Chinese elderly women than in Chinese elderly men.

Keywords: factorial invariance, resilience, aged, factor analysis, reliability

INTRODUCTION

Given China's very large population and the recent sharp increase in the aging population in China, the physical and mental health of the elderly are attracting substantial attention in China. Defined as an individual's ability to cope with adversity and bounce back from difficult experiences (Campbell-Sills and Stein, 2007), resilience has become an important consideration of geriatric mental health because it is key to enabling elderly persons to overcome adverse psychological problems (Connor and Davidson, 2003; Guo et al., 2015). Resilience, which has been shown to not only help reduce morbidity risk, alleviate loneliness, enhance stress-coping ability, and support the maintenance of cognitive and physical functioning of the elderly, may also relieve depressive symptoms associated with stressful life events (Hildon et al., 2010; Lou and Ng, 2012; Fontes and Neri, 2015; Lim et al., 2015; Niu et al., 2016). Thus, it is of great public health significance to study the resilience of the elderly in China.

The 25-item Connor-Davidson Resilience Scale (CD-RISC), which was developed by Connor and Davidson in 2003 to quantify resilience and assess treatment response, is a widely used clinical tool with very good psychometric ratings (Connor and Davidson, 2003; Windle et al., 2011). However, the factor structure of the CD-RISC differs across countries, living environments, and age bands. In a study of 577 healthy adult American participants, exploratory factor analysis revealed a five-factor CD-RISC structure (personal competence, high standards, and tenacity; trust in one's instincts, tolerance of negative affect, and strengthening effects of stress; positive acceptance of change and secure relationships; control; and spiritual influences) (Connor and Davidson, 2003). Meanwhile, in a study of 1,395 community-dwelling American women over 60 years of age, a four-factor structure (personal control and goal orientation; adaptation and tolerance for negative affect; leadership and trust in instincts; and spiritual coping) was obtained (Lamond et al., 2009). In a study of 783 Spanish entrepreneurs operating in the business services sector, a three-factor structure (hardiness; resourcefulness; and optimism) was obtained (Manzano-García and Ayala Calvo, 2013). Likewise, in a study of 246 Turkish earthquake survivors, a three-factor structure (tenacity and personal competence; tolerance of negative affect; and tendency toward spirituality) was observed (Karairmak, 2010). A three-factor structure (tenacity; strength; and optimism) was also obtained with the Chinese version of the CD-RISC in a study of 560 Chinese residents of Guangdong and Beijing (Yu and Zhang, 2007).

Given the various factor structures reported for the 25-item CD-RISC, Campbell-Sills and Stein revised the scale in 2007 into a refined 10-item single-dimension CD-RISC (CD-RISC-10). In a cohort of 1,743 undergraduates, exploratory and confirmatory analyses demonstrated good internal reliability (Cronbach's $\alpha = 0.85$) and construct validity of the CD-RISC-10 (Campbell-Sills and Stein, 2007), indicating that the abridged CD-RISC is a reliable, valid assessment tool, in addition to being easier to apply clinically, relative to the 25-item CD-RISC, owing to its simplicity. The CD-RISC-10 has been translated into several languages, and it has been tested on various populations including Canadian college women, Danish hospital staff, Khmer adolescents, American competitive long-distance runners, French women, Brazilian young people, Spanish nonprofessional caregivers, and low-income African American men, among others (Lopes and Martins, 2011; Scali et al., 2012; Coates et al., 2013; Duong and Hurst, 2016; Gonzalez et al., 2016; Blanco et al., 2017; Lauridsen et al., 2017; Hébert et al., 2018). The Chinese version of the CD-RISC-10 has been reported to be useful for assessing mental resilience quickly in a cohort of Chinese parents of children with cancer (Ye et al., 2017) and was also reported to have good psychometric properties in a study of Wenchuan earthquake survivors (Wang et al., 2010). In addition to having been widely applied, the CD-RISC-10 has also been shown to have good internal consistency, with Cronbach's α values in the range of 0.81–0.95 (Wang et al., 2010; Aloba et al., 2016; Shin et al., 2018).

Some researchers have reported that exposure to trauma in females is associated with a reduced resilience score (Stratta et al., 2013; Hirani et al., 2016). However, due to the lack of data on measurement invariance across genders, we cannot infer the causes of the differences observed because group comparisons require equivalent measurement. To the best of our knowledge, no confirmatory factor analysis study has tested the measurement invariance of the CD-RISC-10 across gender groups in an elderly Chinese cohort.

The current study had four aims. First, we tested the reliability of the CD-RISC-10 in an elderly Chinese study cohort. Second, we examined the model fit of the CD-RISC-10 in a community sample of Chinese elderly. Third, we examined the factorial invariance of the CD-RISC-10 across gender groups. Finally, upon establishment of adequate factorial invariance, we planned to compare resilience scores between men and women.

MATERIALS AND METHODS

Participants and Procedure

This study was conducted in the communities of Beijing, Shandong and Hunan provinces of mainland China. The questionnaires were distributed by well-trained staff to elderly residents aged 60 years and above who came to the community activity center. The staff provided help for participants who had visual impairment, could not read or fill out the questionnaire themselves. The inclusion criteria of this study were: 60 years old and above; agree to participate in this study. The exclusion criteria included: diagnosed with severe mental illness; insufficient cognitive ability to understand the questionnaire; unable to understand mandarin and therefore unable to complete the questionnaire; cannot fill out the questionnaire due to other reasons. A total of 1,284 participants returned questionnaires, but 46 failed to respond to all 10 items. Thus, the final sample included 1,238 (96.4% completion rate). The mean age of the final sample was 71.64 years [standard deviation (SD) = 7.77]. The final sample consisted of 525 men (42%), with a mean age of 72.47 years (SD = 8.09) and 713 women (58%) with a mean age of 71.02 years (SD = 7.46). The study was approved by the ethics committee of Second Xiangya Hospital, Central South University. All participants provided written informed consent at the time of enrollment.

Instrument

The CD-RISC-10, which consists of 10 items, was derived from the original 25-item CD-RISC. It assesses an individual's mental resilience during the past month, such as "Adapt to change" (see the items in the **Appendix**). Respondents rate each item on a 5-point Likert scale from 0 (not true at all) to 4 (true nearly all the time). The item ratings are summed to produce a scale score ranging from 0 to 40, with higher values implying a greater resilience capability. The Chinese version of the CD-RISC-10 employed in this study has been confirmed to have good internal consistency (Cronbach's $\alpha = 0.851–0.910$) and excellent structure validity in Chinese populations (Wang et al., 2010; Ye et al., 2016, 2017).

Data Analysis

Mean values are reported with standard deviations (SDs). Data management was carried out in SPSS 18.0 and confirmatory factor analysis was conducted in Mplus 6.11. Kolmogorov-Smirnov normality testing on item scores showed significant deviation from the normal distribution (all $p < 0.001$, see **Table 1**)

TABLE 1 | Descriptive statistics for each item of the Chinese CD-RISC-10.

Item	Mean	SD	Item factor loading	Skewness	Kurtosis	Kolmogorov-Smirnov *Z*	*p*
1	2.81	0.99	0.75	−0.74	0.27	8.82	0.000
2	2.92	0.93	0.82	−0.73	0.35	8.64	0.000
3	2.52	1.04	0.74	−0.40	−0.40	7.71	0.000
4	2.93	0.92	0.84	−0.78	0.53	8.93	0.000
5	2.87	0.92	0.80	−0.80	0.63	9.53	0.000
6	2.76	0.96	0.84	−0.79	0.51	9.76	0.000
7	2.76	0.95	0.84	−0.67	0.24	9.47	0.000
8	2.78	1.06	0.73	−0.91	0.39	9.83	0.000
9	2.99	0.94	0.84	−0.93	0.73	8.97	0.000
10	2.87	0.95	0.81	−0.83	0.59	9.36	0.000

Note: SD, Standard deviation.

indicating that the data were not normally distributed. Based on the above, the robust maximum likelihood estimator was chosen for data analysis because it, when applied with a mean-adjusted Chi-square (Satorra-Bentler χ^2) statistic and robust standard errors, yields an unbiased goodness-of-fit index that is robust to nonparametric data (Satorra and Bentler, 2001; Wang et al., 2013). The data analysis was conducted in three steps, as delineated below:

In the first step, reliability analysis was conducted. We used Cronbach's α value, McDonald's Omega coefficient, and test-retest reliability coefficient to determine the reliability of the CD-RISC-10.

In the second step, we used confirmatory factor analysis to test the goodness of fit of the single factor structure of the Chinese CD-RISC-10 in the total sample and each gender group. Chi-square (χ^2) and standardized root mean squared residual (SRMR) tests were employed as absolute fit indexes. Because the χ^2 test can be affected by sample size, especially in large samples, we also applied the root mean square error of approximation (RMSEA) as parsimony fit index and applied the Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) as comparative indexes. The following previously established criteria of acceptability were used: SRMR ≤0.08, RMSEA ≤0.08, CFI ≥ 0.90, and TLI ≥ 0.90 (Hu and Bentler, 1999; Brown, 2006; He et al., 2019).

In the third step, multigroup confirmatory factor analysis was undertaken to evaluate the factorial invariance of the CD-RISC-10 across gender groups. The invariance tests were completed for configural invariance (Model 1), metric invariance (Model 2), scalar invariance (Model 3), strict invariance (Model 4), factor variance/covariance invariance (Model 5), and factor latent mean invariance (Model 6) (He et al., 2018). First, we conducted configural invariance tests (without parameter constraints) to evaluate the latent variable structure across gender groups, the results of which served as a baseline model for subsequent tests. Then, metric invariance was tested based on the configural invariance results with factor loading equivalence constraints imposed to ensure similarity of the observed indicators and underlying traits across gender groups. Next, we applied a scalar invariance test in which we constrained both factor loadings and intercepts of variables equally across genders to test for an intergroup difference in the measured intercept based on the result of last step. Subsequently, strict invariance testing was conducted with factor loading, variable intercepts, error variance constraints equally set. Following the measurement equivalence testing, factor variance/covariance invariance and factor latent mean invariance tests were conducted to evaluate the structural invariance of the Chinese CD-RISC-10. We employed the Bayesian information criterion (BIC) and TLI and CFI changes to evaluate invariance across consecutive models. In accordance with published recommendations (Raftery, 1995; Cheung and Rensvold, 2002; Wu et al., 2012; Xiao et al., 2014), a ΔTLI ≤0.010 and a ΔCFI ≤0.010 with a smaller BIC value were considered evidence of invariance. Finally, a nonparametric test, Mann-Whitney U test, was used to compare CD-RISC-10 scores across the gender groups. Because the Kolmogorov-Smirnov normality test showed that the scores of two samples do not conform to the normal distribution, and therefore we conservatively considered that whether the scores of CD-RISC-10 in Chinese elderly men and women conform to the normal distribution remained uncertain.

RESULTS

Descriptive Data and Analyses of Reliability of the 10-Item Connor-Davidson Resilience Scale

Descriptive statistics, including mean scores with SDs, the skewness, and the kurtosis, for each item of the CD-RISC-10 are reported in **Table 1**. The mean scores (SDs) for item 1 through 10 were 2.81 (0.99), 2.92 (0.93), 2.52 (1.04), 2.93 (0.92), 2.87 (0.92), 2.76 (0.96), 2.76 (0.95), 2.78 (1.06), 2.99 (0.94), and 2.87 (0.95). And the skewness values were −0.74, −0.73, −0.40, −0.78, −0.80, −0.79, −0.67, −0.91, −0.93, and − 0.83 for item 1 to 10 while the kurtosis values were 0.27, 0.35, −0.40, 0.53, 0.63, 0.51, 0.24, 0.39, 0.73, and 0.59. According to the skewness and kurtosis values of each item, it can be seen that the mean score of each item presented a negative skewness distribution, and the kurtosis value was close to 0. Overall, the mean (SD) total CD-RISC-10 scores were 27.60 (8.09) for males and 28.68 (7.39) for females. In our study, the Cronbach's α of the CD-RISC-10 was 0.936, the McDonald's Omega coefficient was 0.83, and the test-retest reliability coefficient was 0.665 after 6 months (N = 124).

Confirmatory Factor Analysis

As reported in **Table 2**, we obtained a good fit index for the full sample, the male group, and the female group. Briefly, all TLI, CFI, RMSEA, and SRMR values were > 0.90, >0.90, <0.08, and < 0.08, respectively, indicating that the single-factor model fit the data well in the total sample and each gender group. These results confirmed that the single-factor model can be used as a baseline model for subsequent tests.

Factorial Invariance

The factorial invariance test results, including Satorra-Bentler scaled χ^2 values with degrees of freedom, TLI values and inter-model differences, CFI values and inter-model differences, and BIC values are reported in **Table 3**. The fit indexes of each successive model from Model 1 to Model 4 met the satisfactory fit criteria. That is, between successive models (1 to 2, 2 to 3, and 3 to 4), the ΔTLIs were all <0.010 and the ΔCFIs were all <0.010. The successive decreases in BIC values were 57.891 from Model 1 to Model 2, 53.561 from Model 2 to Model 3, and 60.839 from Model 3 to Model 4. Because these four steps of measurement invariance were performed in sequence, we drew the conclusion that the assumption of measurement invariance across gender was established.

The TLI and CFI values were unchanged from Model 4 to Model 5 (variance/covariance equivalent), with only a 1.816 decrease in BIC. Similarly, from Model 5 to Model 6 (factor latent mean invariance), there was no change in TLI and only a negligible change in CFI, with a BIC decrease of only 0.993. Hence, the ΔTLI and ΔCFI were < 0.010 in both comparisons, with BIC values smaller than in the factor variance/covariance equivalent model. Therefore, we concluded that the factorial invariance across gender among Chinese elders was established.

TABLE 2 | Goodness of fit indexes for the CD-RISC-10 model.

Group	S-Bχ^2	*df*	TLI	CFI	RMSEA	SRMR
Full sample (N = 1,238)	233.195	35	0.949	0.961	0.068	0.031
Males (N = 525)	117.143	35	0.952	0.963	0.067	0.032
Females (N = 713)	145.940	35	0.950	0.961	0.067	0.033

Note: S-Bχ^2, Satorra-Bentler scaled χ^2; df, Degrees of freedom; TLI, Tucker-Lewis Index; CFI, Comparative fit index; RMSEA, Root mean square error of approximation; SRMR, Standardized root mean squared residual.

Gender Difference

The female group had a higher total CD-RISC-10 score, at 28.68, than the male group, at 27.60 (Z = −2.373, p = 0.018). On items 4, 6, 8, and 9, the female group had significantly higher scores than the male group (all p < 0.05). On items 1, 2, 3, 5, 7, and 10, there was no significant difference in scores between the two groups (all p > 0.05). The mean rank and p of each item and the comparison between the two gender groups are given in **Table 4**.

DISCUSSION

The main aim of our study was to probe the psychometric properties of the Chinese version of the CD-RISC-10 in an elderly Chinese population. The Cronbach's α value and the test-retest reliability coefficient indicated that the single-factor Chinese CD-RISC-10 has good internal consistency. The present findings indicate that the CD-RISC-10 is a stable and consistent measurement.

Subsequent multiple group confirmatory analysis performed to estimate the measurement equivalence of the scale across genders showed that the model fitted well in the full sample and in each gender group. Importantly, the results supported configural invariance, metric invariance, scalar invariance, strict invariance, factor variance/covariance invariance, and factor latent mean invariance across genders, confirming full equivalence of the scale across genders. Configural invariance indicates that the pattern of fixed and free parameters was equivalent across genders, with a similar psychological structure being reflected by the same variables in men and women. Subsequent establishment of metric invariance revealed that the relative factor loadings of the items were also equivalent between the two gender groups, indicating that individuals with the same scores on latent variables also scored equally on observation items. In terms of achieving scalar invariance, it was demonstrated that the

TABLE 3 | Fit indices for invariance tests of the CD-RISC-10.

Model	S-Bχ^2	*df*	TLI	CFI	RMSEA (90%CI)	SRMR	Comparison	ΔTLI	ΔCFI	BIC
1	261.409	70	0.951	0.962	0.066 (0.058–0.075)	0.033	–	–	–	26484.037
2	276.489	79	0.955	0.961	0.064 (0.055–0.072)	0.036	2 vs. 1	0.004	–0.001	26426.146
3	295.703	88	0.958	0.959	0.062 (0.054–0.070)	0.036	3 vs. 2	0.003	–0.002	26372.585
4	286.384	98	0.965	0.962	0.056 (0.048–0.063)	0.037	4 vs. 3	0.007	0.003	26311.746
5	289.818	99	0.965	0.962	0.056 (0.048–0.063)	0.064	5 vs. 4	0.000	0.000	26309.930
6	294.466	100	0.965	0.961	0.056 (0.049–0.064)	0.068	6 vs. 5	0.000	–0.001	26308.937

Note: Model 1, Configural invariance; Model 2, Metric invariance; Model 3, Scalar invariance; Model 4, Strict invariance; Model 5, Factor variance/covariances invariance; Model 6, Factor latent mean invariance; S-Bχ^2, Satorra-Bentler scaled χ^2; df, Degrees of freedom; TLI, Tucker-Lewis Index; CFI, Comparative fit index; RMSEA, Root mean square error of approximation; CI, Confidence interval; SRMR, Standardized root mean squared residual; BIC, Bayesian information criterion.

TABLE 4 | Comparisons of each item of CD-RISC-10 between male and female.

Item	Mean rank		*Z*	*p*
	Male	Female		
1	601.39	632.84	−1.164	0.107
2	602.40	632.09	−1.530	0.126
3	601.00	633.12	−1.630	0.103
4	584.97	644.92	−3.101	0.002
5	609.71	626.71	−0.884	0.377
6	597.36	635.81	−1.998	0.046
7	605.59	629.74	−1.251	0.211
8	595.72	637.01	−2.128	0.033
9	585.12	644.81	−3.089	0.002
10	609.37	626.96	−0.910	0.363
Total score	591.44	640.16	−2.373	0.018

observed variable intercepts and CD-RISC-10 reference points were the same for men and women. The attainment of strict invariance suggests that differences in latent variable variation could reflect the observed variable variation differences of the scale. Factor variance/covariance invariance and factor latent mean invariance (a.k.a. structural invariance) were established in the current study, indicating that the observed variables and latent variables possessed the same relationship across the two groups. Consequently, we have concluded that the Chinese version of the CD-RISC-10 estimates latent resilience equivalently across genders and thus can be used to compare mental resilience between elderly men and women in China.

The present finding of a significantly higher CD-RISC-10 total score in women than in men suggests that elderly Chinese women may be generally more resilient than elderly Chinese men, consistent with a previous study in China (Lei et al., 2008). However, in other countries, some studies have reported higher resilience scores for men than women (Stratta et al., 2013). It has been hypothesized that males may be better adapted to traumatic events than women, thus resulting in a "protective model" of resilience (Luthar and Zelazo, 2003). This inconsistency between findings obtained in China and findings obtained elsewhere may be due to social and cultural differences. In China, women are encouraged to seek help, which may yield a stronger social support system for dealing with the pressures of life (Lei et al., 2008). Our findings suggest that elderly women in China may be able to deal with negative emotions, such as stress, more easily and with a faster stress recovery than men in China.

CONCLUSION

The results of this study indicate that the Chinese version of the CD-RISC-10 has good reliability and meets resilience measurement standards well when administered to both elderly men and elderly women living in Chinese communities. Thus, it can be applied as a reliable tool for testing mental resilience and performing inter-gender comparisons of mental resilience. To the best of our knowledge, this study was the first study to assess the factorial invariance of the CD-RISC-10 in elderly Chinese men and women. Our findings confirmed that factorial invariance of the CD-RISC-10 had been established across gender among Chinese elders. Finally, the present results provide evidence of Chinese elderly women having better mental resilience than Chinese elderly men, and further demonstrated that this gender difference could not be attributed to a gender-dependent scale variance, but rather reflect a true gender difference.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

SY conceived and designed the study. LL and JY supervised the study. MM performed the analysis and wrote paper. JH contributed to the analysis. YG and HZ collected the data. All co-authors revised and approved the version to be published.

FUNDING

This research was supported by grants from the National Science and Technology Project for Professional Basic Research (grant number 2015FY111600) and the National Natural Science Foundation (grant number: 81370034).

REFERENCES

Aloba, O., Olabisi, O., and Aloba, T. (2016). The 10-item Connor-Davidson resilience scale: factorial structure, reliability, validity, and correlates among student nurses in southwestern Nigeria. *J. Am. Psychiatr. Nurses Assoc.* 22, 43–51. doi: 10.1177/1078390316629971

Blanco, V., Guisande, M. A., Sánchez, M. T., Otero, P., and Vázquez, F. L. (2017). Spanish validation of the 10-item Connor-Davidson Resilience Scale (CD-RISC 10) with non-professional caregivers. *Aging Ment. Health* 23, 183–188. doi: 10.1080/13607863.2017.1399340

Brown, T. A. (2006). *Confirmatory factor analysis: For applied research*. New York: Guilford Pubn.

Campbell-Sills, L., and Stein, M. B. (2007). Psychometric analysis and refinement of the Connor-Davidson resilience scale (CD-RISC): validation of a 10-item measure of resilience. *J. Trauma. Stress.* 20, 1019–1028. doi: 10.1002/jts.20271

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Coates, E. E., Vicky, P., and Dedrick, R. F. (2013). Psychometric properties of the Connor-Davidson resilience scale 10 among low-income, African American men. *Psychol. Assess.* 25, 1349–1354. doi: 10.1037/a0033434

Connor, K. M., and Davidson, J. R. T. (2003). Development of a new resilience scale: the Connor-Davidson resilience scale (CD-RISC). *Depress. Anxiety* 18, 76–82. doi: 10.1002/da.10113

Duong, C., and Hurst, C. P. (2016). Reliability and validity of the Khmer version of the 10-item Connor-Davidson Resilience Scale (Kh-CD-RISC10) in Cambodian adolescents. *BMC. Res. Notes* 9:297. doi: 10.1186/s13104-016-2099-y

Fontes, A. P., and Neri, A. L. (2015). Resilience in aging: literature review. *Cien. Saude Colet.* 20, 1475–1495. doi: 10.1590/1413-81232015205.00502014

Gonzalez, S. P., Moore, E. W. G., Newton, M., and Galli, N. A. (2016). Validity and reliability of the Connor-Davidson Resilience Scale (CD-RISC) in competitive sport. *Psychol. Sport Exerc.* 23, 31–39. doi: 10.1016/j.psychsport.2015.10.005

Guo, L., Guo, Q., Han, G., Lin, Z., and Liu, K. (2015). Mediation effect of mental resilience between psychological stress and mental health in the community-dwelling elderly. *J. Nurs. Sci.* 30, 60–63. doi: 10.3870/hlxzz.2015.03.060

He, J., Zhong, X., Gao, Y., Xiong, G., and Yao, S. (2019). Psychometric properties of the Chinese version of the Childhood Trauma Questionnaire-Short Form (CTQ-SF) among undergraduates and depressive patients. *Child Abuse Negl.* 91, 102–108. doi: 10.1016/j.chiabu.2019.03.009

He, J., Zhong, X., and Yao, S. (2018). Factor structure of the Geriatric Depression Scale and measurement invariance across gender among Chinese elders. *J. Affect. Disord.* 238, 136–141. doi: 10.1016/j.jad.2018.04.100

Hébert, M., Parent, N., Simard, C., and Laverdière, A. (2018). Validation of the French Canadian version of the brief Connor–Davidson Resilience Scale (CD-RISC 10). *Can. J. Behav. Sci.* 50, 9–16. doi: 10.1037/cbs0000092

Hildon, Z., Montgomery, S. M., Blane, D., Wiggins, R. D., and Netuveli, G. (2010). Examining resilience of quality of life in the face of health-related and psychosocial adversity at older ages: what is "right" about the way we age? *The Gerontologist* 50, 36–47. doi: 10.1093/geront/gnp067

Hirani, S., Lasiuk, G., and Hegadoren, K. (2016). The intersection of gender and resilience. *J. Psychiatr. Ment. Health Nurs.* 23, 455–467. doi: 10.1111/jpm.12313

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Karairmak, O. (2010). Establishing the psychometric qualities of the Connor-Davidson Resilience Scale (CD-RISC) using exploratory and confirmatory factor analysis in a trauma survivor sample. *Psychiatry Res.* 179, 350–356. doi: 10.1016/j.psychres.2009.09.012

Lamond, A. J., Depp, C. A., Allison, M., Langer, R., Reichstadt, J., Moore, D. J., et al. (2009). Measurement and predictors of resilience among community-dwelling older women. *J. Psychiatr. Res.* 43, 148–154. doi: 10.1016/j.jpsychires.2008.03.007

Lauridsen, L. S., Willert, M. V., Eskildsen, A., and Christiansen, D. H. (2017). Cross-cultural adaptation and validation of the Danish 10-item Connor-Davidson Resilience Scale among hospital staff. *Scand. J. Public Health* 45, 1–4. doi: 10.1177/1403494817721056

Lei, W., Chen, X., and Chen, J. (2008). Research on resilience of college students. *Chin. J. Health Psychol.* 16, 155–157. doi: 10.3969/j.issn.1005-1252.2008.02.016

Lim, M. L., Lim, D., Gwee, X., Nyunt, M. S. Z., Kumar, R., and Ng, T. P. (2015). Resilience, stressful life events, and depressive symptomatology among older Chinese adults. *Aging Ment. Health* 19, 1005–1014. doi: 10.1080/13607863.2014.995591

Lopes, V. R., and Martins, M. D. C. F. (2011). Factorial validation and adaptation of the Connor-Davidson Resilience Scale (Cd-Risc-10) for Brazilians. *Revista Psicolog.* 11, 36–50.

Lou, V. W. Q., and Ng, J. W. (2012). Chinese older adults' resilience to the loneliness of living alone: a qualitative study. *Aging Ment. Health* 16, 1039–1046. doi: 10.1080/13607863.2012.692764

Luthar, S. S., and Zelazo, L. B. (2003). "Research on resilience: an integrative review" in *Resilience and vulnerability*. ed. S. S. Luthar (Cambridge, UK: Cambridge University Press), 510–550.

Manzano-García, G., and Ayala Calvo, J. C. (2013). Psychometric properties of Connor-Davidson Resilience Scale in a Spanish sample of entrepreneurs. *Psicothema* 25, 245–251. doi: 10.7334/psicothema2012.183

Niu, G., Sun, X., Tian, Y., Fan, C., and Zhou, Z. (2016). Resilience moderates the relationship between ostracism and depression among Chinese adolescents. *Pers. Individ. Differ.* 99, 77–80. doi: 10.1016/j.paid.2016.04.059

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology* 25, 111–163.

Satorra, A., and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66, 507–514. doi: 10.1007/BF02296192

Scali, J., Gandubert, C., Ritchie, K., Soulier, M., Ancelin, M. L., and Chaudieu, I. (2012). Measuring resilience in adult women using the 10-items Connor-Davidson Resilience Scale (CD-RISC). Role of trauma exposure and anxiety disorders. *PLoS One* 7:e39879. doi: 10.1371/journal.pone.0039879

Shin, G. S., Choi, K. S., Jeong, K. S., Min, Y. S., Ahn, Y. S., and Kim, M. G. (2018). Psychometric properties of the 10-item Conner-Davidson resilience scale on toxic chemical-exposed workers in South Korea. *Ann. Occup. Environ. Med.* 30, 52–58. doi: 10.1186/s40557-018-0265-5

Stratta, P., Capanna, C., Patriarca, S., de Cataldo, S., Bonanni, R. L., Riccardi, I., et al. (2013). Resilience in adolescence: gender differences two years after the earthquake of L'Aquila. *Pers. Individ. Differ.* 54, 327–331. doi: 10.1016/j.paid.2012.09.016

Wang, M., Armour, C., Li, X., Dai, X., Zhu, X., and Yao, S. (2013). The factorial invariance across gender of three well-supported models: further evidence for a five-factor model of posttraumatic stress disorder. *J. Nerv. Ment. Dis.* 201, 145–152. doi: 10.1097/NMD.0b013e31827f627d

Wang, L., Shi, Z., Zhang, Y., and Zhang, Z. (2010). Psychometric properties of the 10-item Connor-Davidson Resilience Scale in Chinese earthquake victims. *Psychiatry Clin. Neurosci.* 64, 499–504. doi: 10.1111/j.1440-1819.2010.02130.x

Windle, G., Bennett, K. M., and Noyes, J. (2011). A methodological review of resilience measurement scales. *Health Qual. Life Outcomes* 9:8. doi: 10.1186/1477-7525-9-8

Wu, W., Lu, Y., Tan, F., Yao, S., Steca, P., Abela, J. R. Z., et al. (2012). Assessing measurement invariance of the children's depression inventory in Chinese and Italian primary school student samples. *Assessment* 19, 506–516. doi: 10.1177/1073191111421286

Xiao, J., Kong, T., Mcwhinnie, C. M., Yao, S., Zhu, X., Zhao, S., et al. (2014). The tripartite model for assessing symptoms of depression and anxiety: psychometric properties of the Chinese Version of the Mood and Anxiety Symptoms Questionnaire in patients with essential hypertension. *J. Cardiovasc. Nurs.* 30, 522–528. doi: 10.1097/JCN.0000000000000193

Ye, Z. J., Qiu, H. Z., Li, P. F., Chen, P., Liang, M. Z., Liu, M. L., et al. (2017). Validation and application of the Chinese version of the 10-item Connor-Davidson Resilience Scale (CD-RISC-10) among parents of children with cancer diagnosis. *Eur. J. Oncol. Nurs.* 27, 36–44. doi: 10.1016/j.ejon.2017.01.004

Ye, Z., Ruan, X., Zeng, Z., Xie, Q., Cheng, M., Peng, C., et al. (2016). Psychometric Properties of 10-item Connor-davidson Resilience Scale among Nursing Students. *J. Nurs.* 23, 9–13. doi: 10.16460/j.issn1008-9969.2016.21.009

Yu, X., and Zhang, J. (2007). Factor analysis and psychometric evaluation of the Connor-Davidson Resilience Scale (CD-RISC) with Chinese people. *Soc. Behav. Personal. Int. J.* 35, 19–30. doi: 10.2224/sbp.2007.35.1.19

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

Items of the CD-RISC-10 (English version)

1. Adapt to change
2. Deal with whatever comes my way
3. See humorous side of things
4. Stress makes me stronger
5. Bounce back after illness or injury
6. Believe I can achieve goals despite obstacles
7. Under pressure I stay focused
8. Not easily discouraged by failure
9. Think of myself as a strong person when facing challenges
10. Able to handle unpleasant feelings

frontiers in Psychology

METHODS
published: 04 June 2019
doi: 10.3389/fpsyg.2019.01139

Fusion Validity: Theory-Based Scale Assessment via Causal Structural Equation Modeling

Leslie A. Hayduk[1]*, *Carole A. Estabrooks*[2] *and Matthias Hoben*[2]

[1] *Department of Sociology, University of Alberta, Edmonton, AB, Canada,* [2] *Faculty of Nursing, University of Alberta, Edmonton, AB, Canada*

Fusion validity assessments employ structural equation models to investigate whether an existing scale functions in accordance with theory. Fusion validity parallels criterion validity by depending on correlations with non-scale variables but differs from criterion validity because it requires at least one theorized effect of the scale, and because both the scale and scaled-items are included in the model. Fusion validity, like construct validity, will be most informative if the scale is embedded in as full a substantive context as theory permits. Appropriate scale functioning in a comprehensive theoretical context greatly enhances a scale's validity. Inappropriate scale functioning questions the scale but the scale's theoretical embedding encourages detailed diagnostic investigations potentially challenging specific items, the procedure used to calculate scale values, or aspects of the theory, but also possibly recommends incorporating additional items into the scale. The scaled items should have survived prior content and methodological assessments but the items may or may not reflect a common factor because items having diverse causal backgrounds can sometimes fuse to form a unidimensional entity. Though items reflecting a common cause can be assessed for fusion validity, we illustrate fusion validity in the more challenging context of a scale comprised of diverse items and embedded in a complicated theory. Specifically we consider the Leadership scale from the Alberta Context Tool with care aides working in Canadian long-term care homes.

Keywords: validity, fusion, scale, structural equation, causal

Edited by:
N. Clayton Silver,
University of Nevada, Las Vegas,
United States

Reviewed by:
Lietta Marie Scott,
Arizona Department of Education,
United States
Cameron Norman McIntosh,
Public Safety Canada, Canada

***Correspondence:**
Leslie A. Hayduk
lhayduk@ualberta.ca

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal
Frontiers in Psychology

Received: *10 December 2018*
Accepted: *30 April 2019*
Published: *04 June 2019*

Citation:
Hayduk LA, Estabrooks CA and Hoben M (2019) Fusion Validity: Theory-Based Scale Assessment via Causal Structural Equation Modeling. Front. Psychol. 10:1139. doi: 10.3389/fpsyg.2019.01139

INTRODUCTION

Scale assessment begins by considering each item's methodology, the respondents' capabilities, and the data gathering procedures (American Educational Research Association, 2014). These fundamental assessments are typically supplemented with evidence of convergent and discriminant validity via factor loadings, factor correlations, and factor score correlations (Brown, 2015). The dependence of factor-based assessments on causal structures is seldom acknowledged, and stands in stark contrast to the causal explicitness accorded typical path models (Duncan, 1975; Heise, 1975; Hayduk, 1987; Bollen, 1989). Combining factor and path structures within programs like LISREL, Mplus, and AMOS encouraged causal understanding of the connections between latent factors and their indicators as well as between different latents (Hayduk and Glaser, 2000a,b; Hayduk et al., 2007; Mulaik, 2010; Hayduk and Littvay, 2012). Including both measurement structure and latent-level structure within a single model makes it possible to investigate what Cronbach and Meehl referred to as construct validity—namely a style of validity assessment grounded in a "nomological network" consisting of an "interlocking system of laws which constitute a theory" where the laws might be "statistical or deterministic" (Cronbach and Meehl, 1955, p. 290). Cronbach and Meehl followed the conventions of their time by replacing cause and causal with synonyms like influences, effects, improves, reflects, results in, and acts on (1955 p. 283–289) but their appeal to "intervening variables" and "specific testable hypotheses" (1955 p. 284, 290) clearly parallel the implications of structural equation models (Hayduk, 1987; Bollen, 1989).

We typically know the full and proximal causal foundations of scale scores because we produce the scale's scores via summing, averaging, weighting, or otherwise combining the values of the items to produce the scale's values. We cause the scale's scores to come into existence by our own, often computer assisted, causal actions. The scale's proximal causal foundations are perfectly known because only the items' recorded values directly determine the scale's values. This causal perfection makes scale scores collinear with the constituent items, and precludes using both the items and scale as data in the same model because the scale scores are seemingly "redundant" with the scale's constitutive items. The fact that the items constitute the full and known proximal causal source of the scale's values does not mean the items' causal sources are known. The values of the items themselves might contain mistakes, inaccuracies, or other features thought of as "error," but the undetermined causal foundations of the items themselves do not disrupt the causal production of scale scores by summing or averaging the items. We know precisely and perfectly how those scale values came into existence because we the researcher summed, averaged, or weighted the items' values to create the scale scores, and presumably we made no mistakes in these calculations. We know the proximal causes of the scale's values (the items) even though we typically do not know the distal causes of the scale's values (the causes of the items). We also do not know whether the world correspondingly melds or fuses the items' values in the same way we fused the items in forming the scale's values.

This article presents a method for simultaneously modeling both a scale and its constituent items by employing fixed/known effects leading from the items to the scale, and embedding this researcher-dictated causal segment within whatever substantive causally-downstream variables match the researcher's theory about how the scale should function if the world similarly fused or melded the items. The scale is modeled as a latent variable having the items as it's known/fixed causal foundations, without requiring that the scale scores appear in the data. The scale is modeled as an effect of the items, and the items' causes are modeled in accordance with the researcher's understanding of the relevant substantive variables—possibly as the items originating in a common factor (reflective indicators), possibly not (formative indicators) (Bollen and Lennox, 1991).

Including both the items and the scale within a single model permits stronger scale validity assessment because the researcher-dictated causal construction of the scale can be checked for consistency with the world's causal control of the items. Fusion validity extends construct validity by incorporating the known research-production of the scale from the items, into the theory surrounding those items—in full acknowledgment that the world may or may not similarly fuse or meld the items into a corresponding causally-produced and causally-effective scale entity. The dependence of both fusion validity and construct validity on theoretical considerations precludes reducing either fusion validity or construct validity to "a single simple coefficient" (Cronbach and Meehl, 1955, p. 300) but this is multiply recompensed by the substantive considerations addressing whether or not the researcher's constructed scale functions in accordance with the theory-expanded understanding of the world's causal actions.

We detail the relevant procedural steps in the next section, and subsequently illustrate the procedure using the Leadership scale from the Alberta Context Tool (ACT) using data collected in the Translating Research in Elder Care (TREC) program (Estabrooks et al., 2009a,b,c, 2011; https://trecresearch.ca). We address technical and more general issues in concluding sections.

METHODS

The Logic Underlying Fusion Validity

Figure 1 presents the model structure required for assessing the fusion validity of a hypothetical scale calculated as the average of three indicator items. The imagined scale's values are calculated as

$$\begin{aligned} Scale &= \frac{Item1 + Item2 + Item3}{3} \\ Scale &= (1/3)\,Item1 + (1/3)\,Item2 + (1/3)\,Item3 \\ Scale &= 0.333 Item1 + 0.333 Item2 + 0.333 Item3. \end{aligned}$$

The 0.333 coefficients are fixed, not estimated, because the researcher averages the items to causally produce the scale's values. Scales created from weighted items would employ the weights as fixed causal coefficients. Either way the equation producing the scale's values contains no "error" variable because the items in the averaging-equation constitute the complete set of immediate causes of the scale's values.

Figure 1 depicts two causes of each item—an item true score variable, and an unlabeled error variable representing the net impact of all unspecified causes of that item. A fixed 1.0 coefficient causally transmits each case's entire item true score into that case's reported value for the corresponding item. Estimation of the items' true score variances and covariances will be explained below. If freed for estimation an item's measurement error variance will often be underidentified, so these variances will often be fixed based on the literature, or via procedures discussed in Hayduk and Littvay (2012), and retrospectively checked. The items' error sources contribute indirectly to the scale scores even though the scale remains fully causally "accounted for" and has no error variable.

Abbreviations: LISREL, Mplus, and AMOS are structural equation modeling programs; TREC, Translating Research into Elder Care; ACT, Alberta Context Tool; CONSORT, Consolidated Standards for Reporting Trials.

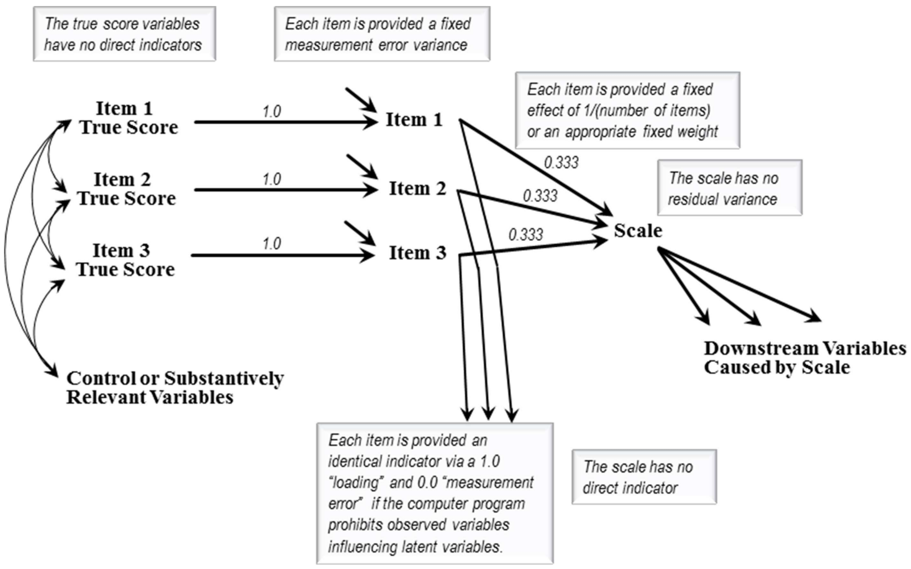

The true score variables have no direct indicators
Each item is provided a fixed measurement error variance
Each item is provided a fixed effect of 1/(number of items) or an appropriate fixed weight
Item 1 True Score
1.0
Item 1
0.333
The scale has no residual variance
Item 2 True Score
1.0
Item 2
0.333
Scale
0.333
Item 3 True Score
1.0
Item 3
Downstream Variables Caused by Scale
Control or Substantively Relevant Variables
Each item is provided an identical indicator via a 1.0 "loading" and 0.0 "measurement error" if the computer program prohibits observed variables influencing latent variables.
The scale has no direct indicator

FIGURE 1 | The basic specification of a fusion validity model.

Assessing fusion validity requires embedding a **Figure 1** style item-and-scale specification into a model containing one or more substantive variables that are causally downstream from the scale, along with whatever control or substantive exogenous variables the researcher specifies. It is the variables causally downstream from the scale that make estimation possible and that potentially underwrite a scale's fusion validity. The fusion in "fusion validity" concerns whether each item fuses (or mixes/combines/merges/melds) with the other items to form a unidimensional scale-entity absorbing and appropriately dispensing the items' causal consequences. That is, a scale displays fusion validity if the items' causal connections to the downstream variables are adequately modeled by the items having fused into a unidimensional variable displaying theorized effects on the downstream variables. If this causal specification fails to match the data, the validity of the scale is questioned, either because the scale is problematic (the fusing is deficient or incomplete) or because the selected downstream variables were ill advised or improperly modeled.

A model requiring additional effects bypassing the scale by leading directly from an item's true scores to a causally downstream variable is reporting the scale's inability to encapsulate that item's effects. The item's effect transmitted though the scale will require enhancement or reduction if the scale's impact on the downstream variable either over- or under-represents the item's impact. No scale-bypassing effects will be required if the items fuse to form a scale capable of functioning as a full and unitary cause carrying the items' effects to the downstream variables. Researchers can certify the immediate causal foundations of the scale because the researcher is in control the scale's construction, but the world will dictate whether the scaled items' causal capabilities correspondingly combine and fuse. The scale—the putatively fused items—and the individual items' true scores constitute potentially contrasting causal explanations for the items' covariances with the downstream variables.

Fusion validity assessment begins with a ***baseline model*** having only the specified items as causes of the scale, and no effects leading directly from the item true scores to any downstream variables (as depicted in **Figure 1**). The scale's validity is supported if this specification fits the data and produces anticipated effect estimates. This baseline model implicitly grants the scale preferential treatment because the scale is permitted effects on the downstream variables while any particular item would have to demand a direct effect by disrupting the baseline model's fit until that item is granted its effect. A model that can only be made consistent with the data by permitting an item to have direct scale-bypassing effects is signaling that the scale is unable to fuse or encapsulate the causal impacts of that item. Scale reassessment is required if an ***amended model*** matches the data after supplementation by scale-bypassing effects but whether the scale should be discarded or usefully-retained depends on the revision details. A model remaining inconsistent with the data even after enhancement by scale-bypassing effects, or other alterations, questions whether the downstream and control variables were sufficiently well-understood to underwrite trustworthy scale assessment.

Examples: Fusion Validity of the Leadership Scale

Our examples employ data from the Translating Research into Elder Care (TREC) archive at the University of Alberta. TREC is a pan-Canadian applied longitudinal (2007-ongoing) health services research program in residential long term care or nursing homes. The TREC umbrella covers multiple ethics-reviewed studies designed to investigate and improve long term nursing-home care (Estabrooks et al., 2009a,c, 2015). We consider the Leadership scale from the Alberta Context Tool which investigates front-line health care aides' perceptions of their care unit work environments. Specifically, we begin with care aide responses to the items comprising the Leadership scale for TREC wave-3 data collected in 2014-2015. The Alberta aides typify the Canadian context by being primarily female (93%), having a first language other than English (61%), and averaging about 46 years of age. We use corresponding Manitoba data to replicate our analysis strategy below, and most Manitoba aides similarly were female (87%), spoke English as a second language (67%), and averaged approximately 45 years of age.

The Leadership scale has undergone traditional measurement assessment (Estabrooks et al., 2009b, 2011) and is calculated by averaging the health care aide's perception of their unit's leader using six 5-point Likert-style items (see **Table 1**). Specifically the Leadership scale is calculated as the average

$$\textit{Leadership Scale} = \frac{\textit{Feedback} + \textit{Success} + \textit{Calmly} + \textit{Listens} + \textit{Mentors} + \textit{Resolves}}{6}$$

which corresponds to

$$\textit{Leadership Scale} = \left(\frac{1}{6}\right)\textit{Feedback} + \left(\frac{1}{6}\right)\textit{Success} + \left(\frac{1}{6}\right)\textit{Calmly} + \left(\frac{1}{6}\right)\textit{Listens} + \left(\frac{1}{6}\right)\textit{Mentors} + \left(\frac{1}{6}\right)\textit{Resolves}.$$

This in turn can be written as an error-free equation containing fixed effect coefficients

$$\textit{Leadership Scale} = (0.167)\,\textit{Feedback} + (0.167)\,\textit{Success} + (0.167)\,\textit{Calmly} + (0.167)\,\textit{Listens} + (0.167)\,\textit{Mentors} + (0.167)\,\textit{Resolves}.$$

Had the scale been defined as a sum or weighted sum, the fixed values in this scale-producing equation would be either 1.0's or the appropriate item weights.

Figure 2 depicts the production of the Leadership scale, along with the effects of Leadership on several interrelated downstream variables. The attitudinal indicators of the downstream variables and the items comprising the scale are each assigned 5% measurement error variance in the models we consider. The exogenous control variables are assigned the following measurement error variances: Sex 1%, Age 5%, English as first language 5%, For-Profit organization 0%, Enough Staff 5%, and Aggressive acts (negative resident behavioral responses) 5%. The leadership items' measurement errors are included at the latent level of the model to correspond to routine construction of scales from error-containing items rather than from item true scores.

Assessing a scale's fusion validity begins with a ***baseline*** model, and may or may not require construction of an ***amended*** model. The baseline model includes:

the items' contributions to the scale,
the scale's effects on the downstream variables,
any effects among the downstream variables,
the control variables' covariances with the scale items
and the control variables' theorized connections to the downstream variables,

but

no direct effects of the items on the downstream variables,
and *no* effects leading directly to the scale (beyond the scale's items).

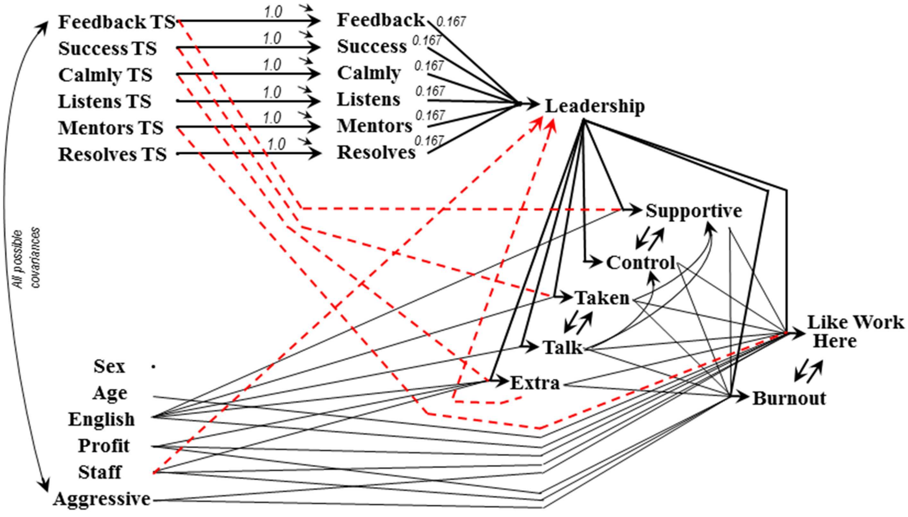
A baseline model displaying clean fit and theory-consistent estimates supports the scale's validity. Item effects bypassing the scale, or additional effects leading to the scale, may appear in an amended model but such effects constitute evidence recommending scale reassessment. Syntax for both the baseline and amended Leadership models is provided near the end of this article.

Both the baseline and amended models might fit or fail to fit, but even a failing baseline model should provide somewhat-reasonable estimates because wild baseline estimates potentially indicate the scale is being encumbered by non-sensical theory-claims about the scale's connections to the downstream variables. Limited modifications to the baseline model are permitted if they maintain the features listed above but such modifications should respect and preserve evidence more appropriately seen

TABLE 1 | Scale items and other variables.

Items	Designation
Leadership scale items	
The degree to which the aide agrees the identified formal leader of their unit:	
Looks for feedback even when it is difficult to hear	Feedback
Focuses on successes rather than failures	Success
Calmly handles stressful situations	Calmly
Actively listens, acknowledges, and then responds to requests and concerns	Listens
Actively mentors or coaches performance of others	Mentors
Effectively resolves conflicts that arise	Resolves
Other variables	
I am a member of a supportive work group	Supportive
I have control over how I do my work	Control
My observations about resident conditions are routinely taken seriously by those in positions of authority	Taken
I am comfortable talking about resident care issues with those in positions of authority	Talk
How often do you have time to do something extra for residents	Extra
In general, I like working here	Like Work Here
I feel burned out from my work	Burnout
Sex	Sex
Age	Age
English first language	English
For-profit organization	Profit
We have enough staff to get necessary work done	Staff
Number out of six possible kinds of resident reactive behaviors experienced in the last 5 shifts	Aggressive

The Leadership scale is the average (mean) of the six Leadership items. The "Other Variables" are single response items, some of which are defined as contributing to scales in other contexts.
Most items are scored 1= strongly disagree, 2 =disagree, 3 = neither agree nor disagree, 4 = agree, 5= strongly agree.
Extra is scored 1=never, 2=rarely, 3=occasionally, 4=frequently, 5=almost always.
Sex: 1= male, 2 = female.
Age: in decade-delimited years.
English: 1 = English first language, 0 = Other first language.
Profit: 1 = working in a for-profit organization, 0 = working in a not-for-profit organization.



Indicator variables and most residual variables have been omitted.
TS = True Score
Dashed red lines are effects added in transforming the baseline model into the enhanced model.

FIGURE 2 | Leadership baseline and enhanced models for Alberta.

TABLE 2 | Model tests.

	χ^2	df	P
Alberta baseline	199.0	67	0.000
Alberta amended	70.5	61	0.189
Manitoba baseline	113.9	68	0.000
Manitoba amended	82.8	66	0.079

χ^2 = chi-square,
df = degrees of freedom,
P = probability.

as questioning the scale's construction. The modifications to the baseline Leadership model for the Alberta data were minimized and fastidiously critiqued (by *LH*) because we planned to subsequently employ the same baseline model with Manitoba data. The objective here was ***not*** to attain fit, but to ensure that the portions of the model concerning the downstream and control variables provided a reasonable theory-context for the Leadership scale. In fact, the resultant Alberta baseline Leadership model remained highly significantly ill fitting (χ^2 = 199.0, *df* = 67, *p* = 0.000, see **Table 2**), suggesting the Leadership scale does not adequately fuse or encapsulate the causal impacts of the leadership items. The baseline model retained all the initially postulated effects whether significant or insignificant. Insignificant estimates constitute unfulfilled theory expectations but they also constitute a cataloged theory-reserve potentially buttressing modifications introduced during construction of an amended model.

Amending a failing baseline model focuses on additional effects emanating from the items and/or effects leading to the scale—namely the effects expressly excluded from the baseline model. Additional item effects will usually originate in the item true-scores because the measurement errors contributing to the observed items are not expected to impact downstream variables. Coefficients suggested by the modification indices were considered individually and added sequentially, based on the *post-hoc* theoretical palatability of their signs, magnitudes, and modeling consequences (such as avoiding underidentification) but for brevity we proceed as if six effects (detailed in the **Appendix** model syntax) were added simultaneously to create the enhanced Leadership model. The amended model fits according to χ^2 with $p = 0.19$ (**Table 2**) and provides the estimates in **Table 3**. The baseline and amended models permit seven possible direct Leadership-scale effects on the downstream variables. All seven estimates were in the anticipated direction, and five were significant, but these effects do not accurately portray the full effectiveness of some of the items on the downstream variables. Four of the six coefficients added in forming the amended model are item effects bypassing the Leadership scale by leading directly from an item's true score to a downstream variable. The effects are: Feedback to Supportive Group, Success to Observations Taken Seriously, Calmly to Time for Something Extra, and Leader Mentors to Like Working Here. These effects lead from four different items' true scores to four different downstream variables and hence cannot be dismissed as artifacts created by a single problematic item.

Each scale-bypassing effect corresponds to an indirect effect transmitted from the item's true score, through the item's observed score, to the scale, and finally to the same downstream variable, as depicted in **Figure 3**. Forming a scale by averaging items forces each item to have the same relatively small indirect effect on any specific downstream variable. For example, for

TABLE 3 | Amended leadership model.

		Supportive	Control	Taken	Talk	Extra	Like Here	Burnout	Leadership	Sex	Age	English	Profit	Staff	Aggressive	TS Feedback	TS Success	TS Calmly	TS Listens	TS Mentors	TS Resolves	R^2
Supportive	AB		−0.388*	0.224*	0.148*				0.473*			0.236*				−0.102*						0.324
	MB		−0.150	0.099*	0.104*				0.499*			0.105										0.219
Control	AB	0.556*			0.090*				0.103													0.331
	MB	0.332*			0.101*				0.257*													0.215
Taken	AB				0.167*				0.455*			0.216*					−0.084*					0.226
	MB				0.169*				0.627*			0.075						−0.115*				0.241
Talk	AB			0.167*					0.191*			−0.145*										0.080
	MB			0.169*					0.129*			−0.110										0.059
Extra	AB								0.412*			0.321*	0.207*	0.242*				−0.206*				0.217
	MB								0.160*			0.435*	0.309*	0.250*								0.168
Like Here	AB	0.115*	0.120*	0.141*	0.032	0.083*		−0.062*	0.123*		0.026*	0.041	−0.047	0.027	0.001					0.082*		0.355
	MB	0.140*	0.064*	0.114	0.030	0.036		−0.062*	0.098		0.026*	0.013	−0.182*	0.069*	−0.005							0.265
Burnout	AB	−0.152*	0.049	−0.079	−0.086	−0.165*	−0.062*		−0.098				0.137	−0.367*	−0.194*							0.153
	MB	−0.377*	−0.086	−0.066	−0.002	−0.164*	−0.062*		0.181				0.088	−0.324*	−0.229*							0.137
Leadership	AB					0.099*								0.173*								1.00
	MB													0.148*								1.00

The fixed 1.0 and 0.167 coefficients leading to and from the items are not shown.

Alberta N = 1610, Manitoba N = 744. Alberta Browne's $\chi^2 = 70.5$, df = 61, p = 0.19. Manitoba Browne's $\chi^2 = 82.8$, df = 66, p = 0.08.

AB, Alberta; MB, Manitoba; TS, True Score.

Coefficients are unstandardized maximum likelihood estimates from LISREL 9.1 (Joreskog and Sorbom, 2016).

Coefficients in highlighted italics were added in forming the amended model, and the −0.150 effect of Control on Supportive in the MB model was fixed at a researcher-assessed value to ensure identification.

**Indicates the coefficient exceeds two standard errors.*

R^2 = Blocked-Error-R^2 (Hayduk, 2006).

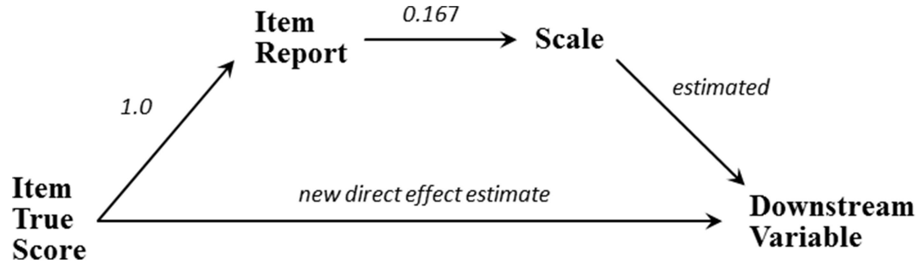


FIGURE 3 | The direct and indirect effects of an item.

TABLE 4 | Effects bypassing the leadership scale in the amended Alberta model.

Effect	Indirect effect of the item via Leadership in the *baseline* model	Indirect effect of the item via Leadership in the *amended* model	Direct effect of the item in the *amended* model	Direct plus indirect effect of the item in the *amended* model
From Feedback to Supportive	0.069	0.079	−0.102	−0.023
From Success to Taken Seriously	0.072	0.076	−0.084	−0.008
From Calmly to Time for Extra	0.036	0.069	−0.206	−0.137
From Mentors to Like Work Here	0.034	0.020	0.082	0.102

The causal variables are the item true scores.
The reported baseline indirect effect = (1.0) (0.167) (estimated scale effect in the Baseline model).
The reported amended indirect effect = (1.0) (0.167) (estimated scale effect in the Amended model).
The direct effects, the indirect effects, and the direct plus indirect effects are "basic effects" (Hayduk, 1987, p. 249) and do not include the enhancements introduced by effects cycling through the loops.

Leadership the indirect effect of the Feedback item on Supportive Group is the product of the 1.0 effect connecting the item's true-score to the observed item, the 0.167 contribution of the item to the Leadership scale, and the scale's estimated 0.473 effect on Supportive Group; which is (1.0)(0.167)(0.473) = 0.079. This indirect effect is identical for all the scale's items because each item's indirect effect begins with 1.0, has the same middle value dictated by the number of averaged items, and employs the same estimated scale-effect on the downstream Supportive Group variable. Thus, each of the six Leadership items has an indirect effect on any specific downstream variable that is one-sixth the Leadership scale's effect on that downstream variable.

An effect leading directly from an item's true score to a downstream variable may either supplement or counteract this indirect effect. An item's total effect is the sum of its direct and indirect effects, so a positive direct effect supplements a positive indirect effect and indicates the item has a stronger impact on the downstream variable than can be accounted for by the scale alone. A negative direct effect counteracts a positive indirect effect and indicates the scale provides an unwarrantedly strong connection between the item and downstream variable. For Leadership three of the four direct effects of items on downstream variables are negative, indicating that requiring these items to work through the Leadership scale produces artificially and inappropriately strong estimates of these items' effects on the applicable downstream variables (**Table 4**). The lone positive direct effect indicates one item (Mentors) should be granted a stronger impact on a downstream variable (Like Working Here) than the Leadership scale permits.

The guaranteed-weak indirect effects of items acting through scales are susceptible to being overshadowed by effects leading directly from the items to downstream variables. All three negative direct item effects in the amended Leadership model, for example, are stronger than the items' small-positive effects carried through the Leadership scale. Two of these direct item effects essentially nullify the corresponding indirect effects, but the third produces a noticeable net negative (reversed) impact (**Table 4**). The Leadership scale's validity is clearly questioned whenever an item's direct effect nullifies or reverses an effect purportedly attributable to the scale containing that item. Direct effects substantially enhancing an item's indirect effect through the scale similarly question the scale (e.g., the direct effect from Mentoring to Like Working Here) because this also signals the scale's inability to appropriately represent the item's causal capabilities. Only four of 42 possible direct effects of the six items on the seven downstream variables are required in the enhanced Leadership model but these effects clearly recommend theoretical reconsideration of the Leadership scale. The involvement of several different scale items and several different outcome variables make the theory challenges somewhat awkward.

The two remaining coefficients added in creating the amended Leadership model lead to the "Leadership scale"—one from an exogenous variable (Have Enough Staff), the other from a downstream variable (Time To Do Something Extra). It is tempting but incorrect to think of these effects as explaining Leadership as originally conceptualized, for example by claiming that health care aides attribute sufficient/insufficient staff to superior/inferior unit leadership as originally scaled. This interpretation is inconsistent with the amended model's estimates because additional causes leading to the scale variable do not explain the original Leadership scale. The new effects redefine the scale such that it only partially corresponds to the original Leadership scale. The original scale was defined as

$$\text{Original Leadership Scale} = (\text{average of six relevant items}).$$

Retaining the same fixed item effects that defined the Leadership scale while adding a new variable's effect changes the equation to

$$\begin{aligned}\text{New Leadership Scale} &= (\text{average of six relevant items})\\ &\quad + (\text{estimated effect of})\ (\text{a newly added cause})\\ \text{New Leadership Scale} &= (\text{Original Leadership Scale})\\ &\quad + (\text{estimated effect of})\ (\text{a newly added cause}).\end{aligned}$$

A predictor variable in an equation does not explain another predictor in that equation, so any additional cause does not explain the original scale, it redefines the scale. The original version of Leadership is transformed into new-Leadership where Enough Staff and Time for Something Extra become components of new-Leadership as opposed to "explaining" anything about Leadership as originally specified and defined. Explaining original Leadership would require explaining the items averaged to create the original Leadership scale.

The downstream variables will usually be included in the model because they are directly caused by the scale, so enhancing a model by adding an effect leading from a downstream variable back to the scale is likely to introduce a causal loop. The additional effect leading from Time for Something Extra to New-Leadership entangles New-Leadership in just such a loop (see **Figure 2**). Though somewhat unusual, causal loops are understandable and not particularly statistically problematic (Hayduk, 1987 Chapter 8; Hayduk, 1996 Chapter 3). A more fundamental concern is that even this single causal loop ensnares Leadership in a causal web that renders it impossible to define or measure Leadership without modeling the appropriate looped causal structure. A variable that was formerly an effect of Leadership becomes both a cause and effect of New-Leadership—and that new causal embeddedness renders standard measurement procedures inappropriate. Items that act as causes can be averaged to create scale scores but we currently have no way of creating scores for "scale" variables trapped in causal loops containing both their causes and effects. The only appropriate option is to place a "scale" like New-Leadership in a model respecting the relevant causal complexities. That stymies traditional scale score calculations even though it employs the same observed variables and permits valid investigation of the causal connections between the scale items, the scale, and the downstream variables.

We now briefly consider the fusion validity of the Leadership scale using data from health care aides in the Canadian province of Manitoba. The Manitoba model employs the same percentage of measurement error variance as in Alberta and is structured identically to the baseline Alberta model with the exception that the smaller of one pair of downstream reciprocal effects was provided a small fixed value (Supportive to Control, −0.150) to avoid underidentification—which results in the baseline Manitoba model having one more degree of freedom than the Alberta model. The Manitoba baseline Leadership model, like the Alberta baseline model, was highly significantly inconsistent with the data (**Table 2**). Amending the model by freeing one item's effect on a downstream variable (Calmly Handles to Observations Taken Seriously) and permitting the exogenous variable Enough Staff to influence "Leadership" resulted in a model that fit nearly as well as the amended Alberta model and with similar estimates (**Tables 2**, **3**).

The small number of demanded alterations is comforting but the repeated requirement for an effect of the control variable Enough Staff on "Leadership" is particularly noteworthy. Two separate data sets report that "Leadership" as perceived by health care aides should be redefined to include Enough Staff in order to make the Leadership scale consistent with the evidence. The remaining alterations differ between the Alberta and Manitoba models, including the challenging loop-creating effect, and these clearly warrant additional investigation. But rather than pursuing the substantive details of these Leadership models, we turn to more general technicalities involved in assessing fusion validity.

Technicalities, Extensions, and Potential Complexities

We developed fusion validity to investigate scales developed by researchers participating in TREC (Translating Research into Elder Care) studies of residents and care aides in long-term care facilities (Estabrooks et al., 2009a) and not as an intentional continuation or extension of specific statistical traditions. We thank one of our reviewers for encouraging us to report and reference connections between fusion validity and various threads within the statistical and methodological literature. Fusion validity's grounding in causal networks places it closer to the causal-formative (rather than composite-formative) indicators discussed by Bollen and Bauldry (2011), and fusion validity's dependence on context-dependent theory distances it from some components of traditional classical test theory. The inclusion of both a scale and its items within the same model provides an opportunity to reassess the points of friction evident in exchanges between Hardin (2017) and Bollen and Diamantopoulos (2017). The points are too diverse and complex for us to resolve, though we hope our comments below provide helpful direction.

Fusion validity's dependence on embedding the scale in an appropriate causal context raises potential technical as well as theoretical concerns. The baseline model may fit, or fail to fit, and either result may prove problematic. A fitting baseline model containing unreasonable estimates questions whether the control and downstream variables are sufficiently well-understood to be entrusted with scale adjudication. Nothing forbids a few mild modifications to initially-failing baseline models but it may be technically tricky to avoid inserting coefficients more appropriately regarded as scale-confronting. Reasonable modifications might rectify downstream variables' causal interconnections, or exogenous control variables' connections to the downstream variables, but ferreting out whether or not a modification questions the scale may prove difficult. For example, if a control variable correlates substantially with an item's true-scores the modification indices may equivocate between whether the control variable or the item effects a downstream variable, and thereby equivocate between whether the researcher is confronting scale-compatible or scale-incompatible evidence. Baseline models having complicated interconnections among the downstream variables, or unresolved issues with multiple indicators of control or downstream variables are likely to prove particularly challenging. Neophytes may have difficulty recognizing, let alone resisting, coefficients that could lead to inappropriately obtained model fit, especially knowing that persistent baseline model failure questions their scale. Validity requires consistency with our understandings, but when our modeled understandings (whether in a baseline or amended model) are problematic, concern for validity transmutes into concern for the fundamental commitments underlying scientific research.

Standardized residual covariances typically provide diagnostic direction, but they provided minimal assistance in fusion validity assessments because the scale latent variable and the item true-score latents have no direct indicators and consequently contribute only indirectly to the covariance residuals. Furthermore, the residual covariance ill fit among the scale items should be essentially zero because the model's structure nearly guarantees that the estimated covariances among the item true scores should reproduce the observed item covariances irrespective of the number or nature of the items' sources. This "guaranteed" perfect fit among the items might be thought of as a diagnostic limitation, but it is more appropriately thought of as convincingly demonstrating that fusion validity does not depend on the items having a common factor cause. The free covariances among the item true scores permit the items to reflect a single factor, but also permit the item true scores to reflect multiple different "factors." Thus, fusion validity can assess scales created from both reflective and formative indicators (Bollen and Lennox, 1991). The issue addressed by fusion validity is not the source of the items but whether the items causally combine into a scale that is unidimensional in its production of downstream variables. Fusion validity is not about the dimensionality of the scale variable. The scale variable is unavoidably unidimensional no matter the number of constituent items or the number of "factors" producing those items. The issue is the *causal fidelity of fusing the potentially-diverse items* into a unidimensional variable capable of transmitting the potentially-diverse items' effects to the downstream variables.

If the baseline model fails after exhausting reasonable modifications, the focus switches to scale-questioning connections between specific items and the downstream variables, and/or additional effects leading to the scale in an amended model. Here the most useful diagnostics are the modification indices and expected parameter change statistics. A large, not merely marginally-significant, modification index for an item's effect on a downstream variable, combined with an implicationally-understandable expected parameter change statistic, would suggest including a coefficient speaking against the scale. The magnitude and sign of the expected parameter change statistic for an item's direct effect should be understandable in the context of the indirect effect that the item transmits through the scale as discussed in regard to **Figure 3**. A scale-bypassing effect speaks against the thoroughness of the encapsulation provided by the scale but if the world contains multiple indirect effect mechanisms (Albert et al., 2018), it might require both a direct item effect and the indirect effect acting through a fused scale. Unreasonably-signed scale bypassing effects speak more clearly against the scale.

If one specific item requires stronger (or weaker) effects on multiple downstream variables, and if the required effect adjustments are nearly proportional to the scale's effects, that might be accommodated by strengthening (or weakening) the item's fixed effect on the scale. For example, a substantial modification index corresponding to one item's fixed 0.167 effect leading to the Leadership scale might recommend constructing a weighted Leadership scale rather than the current average scale. Similarly, if the baseline model contained fixed unequal item weightings, large modification indices for some weights might recommend reweighting the items.

It should be clear that an amended model requiring a direct effect of an item's true-score on a downstream variable is not equivalent to, and should not be described as, having altered the item's contributions via the scale. Effects transmitted via the scale must spread proportionately to all the variables downstream from the scale. An effect leading from one item to a specific downstream variable disrupts the scale's proportional distribution requirement for that specific pairing of an item and downstream variable. The new direct effect also loosens ("partially frees") the constraints on that item's effects via the scale on the other downstream variables because these other effects need no longer be rigidly proportional to this item's effect via the scale on the bypass-receiving downstream variable. The proportionality constraints on the other items' effects (via the scale) on the downstream variables are also slightly loosened by the scale-bypassing effect but the greater the number of items and scale-affected downstream variables the feebler the loosening of these constraints. Each additional scale-bypassing effect progressively, even if minimally, loosens the proportionality constraints on all the items' effects on the downstream variables via the scale. This suggests an accumulation of minor constraint relaxations resulting from multiple scale-bypassing effects in an amended model might constitute holistic scale-misrepresentation.

A substantial modification index might also be connected to the fixed zero variance assigned to the residual variable that causes the scale—namely the zero resulting from the absence of an error variable in the item-averaging equation constructing the scale. A substantial modification index here suggests some currently unidentified variable may be fusing with the modeled scale items, or that there are some other unmodeled common causes of the downstream variables. A scale known to be incomplete due to unavailability of some specific cause might warrant assigning the scale's residual variance a fixed nonzero value, or possibly a constrained value. The scale's residual variance might even be freed if sufficient downstream variables were available to permit estimation. A nonzero residual variance should prompt careful consideration of the missed-variable's identity. The potential freeing of the scale's residual variance clearly differentiates fusion validity from confirmatory composite analysis, which by definition forbids each composite from receiving effects from anything other than a specified set of indicators (Schuberth et al., 2018, p. 3). Indeed, the potential freeing of the scale's residual variance pinpoints a causal conundrum in confirmatory composite analysis—namely how to account for the covariance-parameters connecting composites without introducing any additional effects leading to any composite (Schuberth et al., 2018, Figure 5). This is rendered a non-issue by fusion validity's causal epistemological foundation. The relevant modeling alternatives will be context-specific but likely of substantial theoretical and academic interest.

The fixed measurement error variances on the observed items might also require modification but the implications of erroneous values of this kind are likely to be difficult to detect, and could probably be more effectively investigated by checking the model's sensitivity to alternative fixed measurement error variance specifications. Modeling the items' and/or scale's residual variables as independent latent variables (Hayduk, 1987, p.191-198) would provide modification indices permitting assessment of potential measurement error covariances paralleling the proposals of Raykov et al. (2017). Attending to modification indices, or moving to a Bayesian mode of assessment, would implicitly sidle toward exploration, which nibbles at the edges of validity, so especially-cautious and muted interpretations would likely be advisable.

Other technicalities might arise because the scale variable and the item true score variables have no direct indicators, which forces the related model estimates to depend on indirect causal connections to the observed indicators. The scale's effects on the downstream variables, for example, are driven by the observed covariances between the items' indicators and the indicators of the downstream variables because the scale's effects provide the primary (even if indirect) causal connections between these sets of observed indicators. And the covariances among the "indicatorless" item true scores will mirror the covariances of the observed item indicators because the true scores' covariances constitute the primary causal sources of these covariances. The absence of direct latent to indicator connections may produce program-specific difficulties, as when the indicatorless item true score latents stymied LISREL's attempts to provide start values for these covariances (Joreskog and Sorbom, 2016). This particular technicality is easily circumvented by providing initial estimates approximating the corresponding items' observed variances and covariances.

Related complexities may arise because programs like LISREL require modeling the observed items as perfectly measured latents (with $\lambda = 1.0$, and $\Theta\varepsilon = 0.0$) as in **Figure 1**, which moves the measurement error variances into LISREL's Ψ matrix and places zero variances in $\Theta\varepsilon$, thereby producing an expected and ignorable warning that $\Theta\varepsilon$ is not positive definite. This statistical annoyance arises because the measurement error variance in each item unavoidably contributes to the scale. This could be transformed into an interesting theoretical issue by considering that in some contexts it might be reasonable to think of this as "specific variance" which could be split into an item's measurement error variance dead-ending in the indicator (namely a non-zero $\Theta\varepsilon$ in LISREL) and another part indirectly contributing to the scale and downstream variables (as in the illustrated fixed Ψ specification). In the extreme, a fusion validity model might specify all the item measurement error variance as dead-ending in the indicators so the scale is created from fixed effects arriving from the items' true-scores. This would correspond to moving the fixed effects currently leading to the scale from the observed-items to the true-score items in **Figure 1**, and would permit investigating how a scale would function if it was purified of indicator measurement errors. This version of the fusion validity model would attain the epitome of scale construction—a scale freed from measurement errors—which is unattainable in contexts employing actual error-containing items. Contrasting the behavior of the "measurement error free" and "real" scales would permit assessing whether the unavoidable incorporation of items' measurement errors in the "real" scale introduces consequential scale degradation or interference.

It would be possible to simultaneously assess the fusion validity of two or more different scales constructed from a single set of items if the model contains downstream variables differentially responding to those scales. This opens an avenue for assessing Bollen and Bauldry (2011) differentiation between "covariates" and measures, and it provides a route to resolving the confusions plaguing formative indicators, partial least squares, and item parcels (Little et al., 2013; Marsh et al., 2013; Henseler et al., 2014; McIntosh et al., 2014). Importantly, factor score indeterminacy does not hinder fusion validity assessments. Indeed, if the items were modeled as being caused by a common factor (rather than as having separate latent causes as illustrated), fusion-validity modeling of the scale would provide a potentially informative estimate of the correlation between the factor and the scale (now factor scores).

We should also note that fusion validity surpasses composite invariance testing (Henseler et al., 2016): because fusion validity assessment is possible with a single group, because it employs as sophisticated a theory as the researcher can muster, and because validity supersedes mere reliability/invariance. Introducing a longitudinal component to a fusion validity model would even permit differentiating "specificity" from "error" (Raykov and Marcoulides, 2016a) if the fusion validity model incorporates factor structuring of the items. In general, replacing items with parcels disrupts the item-level diagnostics potentially refining fusion validity models, and hence is not advised. A reviewer noted that attention to non-linearities might "introduce more flexibility (and fun)" into fusion validity. We agree—but quite likely "fun" for only the mathematically-inclined (Song et al., 2013).

Fusion validity's theory-emphasis does not end with formulation of appropriate baseline and amended models—it may extend into the future via consideration of what should be done next. For example, one author (*CE*) was concerned that the demand for parsimony during data collection resulted in omission of causes of leadership, and she was uneasy about employing downstream latents having single indicators instead of similarly named scales having multiple indicators. These seemingly methodological concerns transform into theory-options as one considers exactly how a supposedly-missed cause should be incorporated in an alternative baseline model—namely is the missed variable a control variable, a downstream variable, or possibly an instantiation of the scale's residual variable? These have very different theoretical and methodological implications. Similar detailed theoretical concerns arise from considering how an additional-scale, or multiple indicators used by others as a scale, should be modeled by a researcher investigating a focal scale such as Leadership. Fusion validity models are unlikely to provide definitive-finales for their focal scales but rather are likely to stand as comparative structural benchmarks highlighting precise and constructible theoretical alternatives. An advance in theory-precision is likely, irrespective of the focal scale's fate.

DISCUSSION AND CONCLUSIONS

A scale's fusion validity is assessed by simultaneously modeling the scale and its constituent items in the context of appropriate theory-based variables. Fusion validity presumes the items were previously assessed for sufficient variance, appropriate wordings, etcetera, and that a specific scale-producing procedure exists or has been proposed (whether summing, averaging, factor score weightings, or conjecture). This makes the scale's proximal causal foundations known because the researcher knows how they produce, or anticipate producing, scale values from the items, but whether the resultant scale corresponds to a unidimensional world variable appropriately fusing and subsequently dispensing the items' effects to downstream variables awaits fusion validity assessment.

Fusion validity circumvents the data collinearity between a scale and its constituent items by employing only the items as data while incorporating the scale as a latent variable known through its causal foundations and consequences. The scale is modeled as encapsulating and fusing the items, and as subsequently indirectly transmitting the items' impacts to the downstream variables. An item effect bypassing the scale by running directly to a downstream variable signals the scale's inability to appropriately encapsulate that item's causal powers.

The fixed effects leading from the items to the scale are dictated by the item averaging, summing, or weighting employed in calculating the scale's values. The effects leading from the scale to the downstream variables are unashamedly, even proudly, theory-based because validity depends upon consistency with current theoretical understandings (Cronbach and Meehl, 1955; Hubley and Zumbo, 1996; American Educational Research Association, 2014). After reviewing scale assessments in multiple areas, Zumbo and Chan observed that "by and large, validation studies are not guided by any theoretical orientation, validity perspectives or, if you will, validity theory" (Zumbo and Chan, 2014, p. 323). The unavoidable collinearity between item and scale data ostensibly hindered checking the synchronization between items, scales, and theory-recommended variables—a hindrance overcome by the fusion validity model specification presented here.

It is clear how items caused by a single underlying factor might fuse into a unidimensional scale. The consistent true-score components of the items accumulate and concentrate the underlying causal factor's value while random measurement errors in the items tend to cancel one another out. The simplicity and persuasiveness of this argument switched the historical focus of scale validity assessments toward the factor structuring of the causal source of the items and away from the assessment of whether some items fuse to form a scale entity. Fusion validity examines whether the items fuse to form a unitary variable irrespective of whether or not the items originate from a common causal factor. That is, fusion validity acknowledges that the world's causal forces may funnel and combine the effects of items even if those items do not share a common cause. It is possible for non-redundant items failing to satisfy a factor model to nonetheless combine into a unidimensional scale displaying fusion validity. For example, the magnitude of gravitational, mechanical, and frictional forces do not have a common factor cause, yet these forces combine in producing the movement of objects. The causal world might similarly combine diverse psychological or social attributes into unidimensional entities such as Leadership ability, or the like. Given that diversity among the items' causes does not dictate whether or not those items fuse, it remains possible for items failing to comply with a factor model to nonetheless fuse into valid scales—though the fusing is "not guaranteed" and requires validation.

And the reverse is also possible. Items having a common cause and satisfying the factor model may, or may not, fuse into valid scales. That is, items sharing a common cause do not necessarily have common effects. For example, the number of sunspots is a "latent factor" that causes both the intensity of the northern lights and the extent of disruption to electronic communications but we know of no causally downstream variable responding to a fused combination of northern light intensity and communication disruption. In brief, fusion validity focuses on whether the items' effects combine, meld, or fuse into an effective unidimensional scale entity irrespective of the nature of the items' causal foundations. If a researcher believes their items share a common factor cause and also fuse into a scale dimension, it is easy to replace the item true-score segment of the fusion validity model with a causal factor structure. Such a factor-plus-fusion model introduces additional model constraints and is more restrictive than the illustrated fusion validity model specification. The appropriateness of the additional factor-structure constraints could be tested via nested-model χ^2-difference testing, and might be informative, but would not be required for fusion validity. Fusion validity can therefore be applied to both reflective and formative indicators.

Evidence confronting a scale arises when a failing baseline model must be amended: by introducing item effects bypassing the scale on the way to downstream variables, by introducing additional effects leading to the scale, by altering the fixed effects constituting the scale's calculation, or by altering the error variance specifications. An effect leading directly from an item to a downstream variable alters the understanding of the scale irrespective of whether that effect supplements or counteracts the item's indirect effect through the scale. Either way, the scale is demonstrated as being incapable of appropriately encapsulating the item's causal consequences, and hence retaining both the item and scale may be required for a proper causal understanding. An item effect bypassing the scale does not necessarily devastate the scale because it is possible for several items to fuse into an appropriate scale entity having real effects and yet require supplementation by individual item effects. Items having direct effects on downstream variables that cancel out or radically alter the item's indirect effect via the scale are more scale-confronting. Scale-bypassing effects and other model modifications encourage additional theory precision—precision which is likely to constitute both the most challenging and the most potentially-beneficial aspect of fusion validity assessment.

Amending the baseline model by introducing an additional effect leading to the scale variable—namely an effect beyond the originally scale-defining item effects—produces a new and somewhat different, but potentially correct, scale variable. The new effect does not explain the original scale. Both the original scale and new-scale are fully explained because both scales typically have zero residual error variance. They are just different fully explained variables which possess and transmit somewhat different effects. The new scale variable may retain the ability to absorb and transmit the original items' effects to the downstream variables but the new scale is also capable of absorbing and transmitting the actions of the additional causal variable. The researcher's theory should reflect a scale's changing identity. Both theory and methods are likely to be challenged by attempting to expunge the old scale scores from the literature—especially since the new scale's scores would not be calculable in existing data sets lacking the new scale-defining variable.

Both theory and methods are likely to be more strongly challenged if model alteration requires effects leading to the scale from downstream variables because such effects are likely to introduce causal loops. Loops provide substantial, though surmountable, theory challenges (Hayduk, 1987, 1996, 2006; Hayduk et al., 2007) but they introduce especially difficult methodological complications because there is no standard procedure for obtaining values for scales entangled in loops containing their effects. A model can contain as many equations as are required to properly model looped causal actions but the single equation required for calculating a scale's scores becomes unavoidably misspecified if the equation contains one of the scale's effects as a contributory component. If a substantial modification index calls for a loop-producing effect that effect would likely be identified. In contrast, theory-proposed looped effects may prove more difficult to identify (Nagase and Kano, 2017; Wang et al., 2018; Forre and Mooij, 2019).

The requirement that valid scales function causally appropriately when embedded in relevant theoretical contexts implicitly challenges factor models for having insufficient latent-level structure to endorse scale validity. Indeed, fusion validity assessment supersedes numerous factor analytic "traditions." The lax model testing evident in even recent factor analysis texts contrasts with the careful testing required for the baseline and enhanced fusion validity models (Hayduk, 2014a,b; Brown, 2015). And if a baseline or enhanced model is inconsistent with the downstream variables, researchers steeped in traditional factor practices are likely to reflexively attempt to "fix" the model by inserting indicator error covariances or by deleting indicators, rather than retaining the indicators and adding theory-extending latents. Adding latents implicitly challenges the multiple indictors touted by factor analysis because adding latents while retaining the same indicators sidles toward single indicators (Hayduk and Littvay, 2012). Researchers from factor analytic backgrounds are likely to find it comparatively easy to sharpen their model testing skills but will probably encounter greater difficulty pursuing theoretical alternatives involving effects among additional similar latent variables, or appreciating how items having diverse causal backgrounds might nonetheless combine into an effective unidimensional causal entity—such as leadership, trust, stress, or happiness. The tight coordination between theory and scale validity assessment provides another illustration of why measurement should accompany, not precede, theoretical considerations (Cronbach and Meehl, 1955; Hayduk and Glaser, 2000a, Hayduk and Glaser, 2000b).

Scales were traditionally justified as more reliable than single indicators, and as easier to manage than a slew of indicators. Both these justifications crumble however, if the scale's structure is importantly causally misspecified, because invalidity undermines reliability, and because a causal-muddle of indicators cannot be managed rationally. In medical contexts, for example, it is unacceptable to report a medical trial's outcome based on a problematic criterion scale, but equally unacceptable to throw away the data and pretend the scale-based trial never happened. This dilemma underpins the call for CONSORT (the Consolidated Standards for Reporting Trials) to instruct researchers on how to proceed if a scale registered as a medical trial's criterion measure is found to misbehave (Downey et al., 2016). The impact of some assumption violations on scale reliability have been addressed for factor-structured models (Raykov and Marcoulides, 2016b) but if the causal world is not factor structured, the nature and utility of "reliability" remains obscure. And what constitutes "criterion validity" (Raykov et al., 2016) if both the criterion and the scale happen to be involved in a causal loop? Ultimately, avoiding iatrogenic consequences requires a proper causal, not merely correlational, understanding of the connections linking the items, the scale, the downstream variables, and even the control variables. Pearl and Mackenzie (2018) and Pearl (2000) present clear and systematic introductions to thinking about causal structures and why control variables deserve consideration. One of our reviewers pointed us toward a special issue of the journal *Measurement* focused on causal indicators and issues potentially relating to fusion validity. We disagree with enough points in both the target article by Aguirre-Urreta et al.'s (2016) and the appended commentaries that we recommend these exchanges as a practice-exam for anyone considering investigating a fusion validity model. Try to follow the consequences of the Aguirre-Urreta et al. (2016) simulation having: (a) employed causal indicators that do not require any control variables, and (b) having used causal indicators that are forbidden effects bypassing the scale variable. It should also prove instructive to notice the emergent focus on measurement's connection to substantive theory—and not just measurement traditions.

The assessment of fusion validity illustrated above slightly favors the scale by initially modeling the scale's presumed effects, and by permitting baseline model modifications which potentially, even if inadvertently, assist the scale. A scale-unfriendly approach might begin with a baseline model permitting some scale-bypassing item effects, while excluding all the scale's effects on the downstream variables until specific scale effects are demanded by the data. However done, models assessing whether a set of items fuse to form a scale will depend on theory, will focus attention on theory, and will provide opportunities to correct problematic theoretical commitments.

Fusion validity shares traditional concerns for item face validity and methodology but requires variables beyond the items included in the scale—specifically variables causally downstream from the postulated scale but possibly control variables which may be upstream of the items. Fusion validity permits but does not require that the scaled items have a common factor cause, or even that the items correlate with one another.

Traditional formulations make reliability a prerequisite for validity but some forms of reliability are not a prerequisite for fusion validity because fusion validity does not share a factor-model basis. It does require that the items fuse or meld in forming the scale according to the researcher's specifications. Consequently, just as construct validity cannot "be expressed in the form of a single simple coefficient" (Cronbach and Meehl, 1955, p. 300), fusion validity assessment does not produce one single coefficient's value and instead depends on the researcher's facility with structural equation modeling to assess the scale's coordination with whatever substantive variables are required by theory. This means the researcher must be as attentive to the possibility of faulty theory as to faulty scaling—which seems to be an unavoidable concomitant of the strong appeal to theory required by seeking validity. Fusion validity's inclusion in the model of theory-based variables along with both the items and scale permits many assessments unavailable to traditional analyses, and potentially recommends correspondingly diverse theory, scale, and item improvements. Complexity abounds, so only those strong in both their theory and structural equation modeling need apply.

Embedding a scale in deficient theory will highlight the deficiencies, while embedding a scale in trustworthy theory will provide unparalleled validity assessments. Fusion validity assessment does not guarantee progress but provides a way to investigate whether our scales coordinate with our causal understandings, and a way to check whether traditional scale assessments have served us well.

AVAILABILITY OF DATA AND MATERIALS

The data analyzed in this study are from care aides in Alberta and Manitoba collected in 2014-2015 and are archived by the Translating Research into Elder Care (TREC) team at the University of Alberta. TREC is a pan-Canadian applied longitudinal (2007-ongoing) health services research program in residential long term care. The TREC umbrella covers multiple ethics-reviewed studies designed to investigate and improve long term care. The appended LISREL syntax contains the covariance data matrix sufficient for replicating the Alberta estimates or estimating alternative models.

ETHICS STATEMENT

Ethics approval was obtained by the Translating Research in Elder Care team from both universities and all the institutions and participants participating in the reported studies.

AUTHOR CONTRIBUTIONS

LH conceived the analytical procedure, conducted the analyses, wrote the draft article, and revised the article incorporating coauthor suggestions. CE and MH critically assessed the article and suggested revisions. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

Funding was provided by the Canadian Institutes of Health Research (CIHR) and partners in the Ministries of Health in British Columbia, Alberta, and Manitoba, as well as, regional health authorities in participating BC and AB regions.

ACKNOWLEDGMENTS

The authors thank Greta Cummings, Elizabeth Anderson, and Genevieve Thompson for suggesting the downstream variables that should be used; Mike Gillespie for thought-provoking discussions; and Joseph Akinlawon and Ferenc Toth for data and archive assistance.

The authors acknowledge the Translating Research in Elder Care (TREC) 2.0 team for its contributions to this study. As of March 29th, 2017, the TREC 2.0 Team is comprised of the following co-investigators, decision makers, and collaborators listed here in alphabetical order:

Principal Investigator: CE.

Co-investigators: Elizabeth Andersen, Ruth Anderson, Jennifer Baumbusch, Anne-Marie Boström, Whitney Berta, Fiona Clement, Lisa Cranley, Greta G. Cummings, James Dearing, Malcolm Doupe, Liane Ginsburg, Zahra Goodarzi, Andrea Gruneir, LH, Jayna Holroyd-Leduc, Janice Keefe, Jennifer Knopp-Sihota, Holly Lanham, Margaret McGregor, Peter Norton, Simon Palfreyman, Joanne Profetto-McGrath, Colin Reid, Sentil Senthilselvan, Malcolm Smith, Janet Squires, Gary Teare, Genevieve Thompson, Johan Thor, Adrian Wagg, Lori Weeks.

Decision Makers: Carol Anderson, Heather Cook, Laura Choroszewski, Heather Davidson, Lorraine Dacombe Dewar, Roxie Eyer, Hana Forbes, Heather Hanson, Cindy Kozak-Campbell, Barbra Lemarquand-Unich, Keith McBain, Cindy Regier, Irene Sheppard, Corinne Schalm, Deanne (Dee) Taylor, Gina Trinidad.

Consultants: Jeff Poss, Michael Murray.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01139/full#supplementary-material

REFERENCES

Aguirre-Urreta, M. I., Ronkko, M., and Marakas, G. M. (2016). Omission of causal indicators: consequences and implications for measurement. *Measurement* 14, 75–97. doi: 10.1080/15366367.2016.1205935

Albert, J. M., Cho, J. I., Liu, Y., and Nelson, S. (2018). Generalized causal mediation and path analysis: Extensions and practical considerations. *Statist. Methods Med. Res.* 1:962280218776483. doi: 10.1177/0962280218776483

American Educational Research Association (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley and Sons.

Bollen, K. A., and Bauldry, S. (2011). Three Cs in measurement models: causal indicators, composite indicators, and covariates. *Psychol. Methods* 16, 265–284. doi: 10.1037/a0024448

Bollen, K. A., and Diamantopoulos, A. (2017). Notes on measurement theory for causal-formative indicators: a reply to Hardin. *Psychol. Methods* 23, 605–608. doi: 10.1037/met0000149

Bollen, K. A., and Lennox, R. (1991). Conventional wisdom on measurement: a structural equation perspective. *Psychol. Bull.* 110, 305–314. doi: 10.1037/0033-2909.110.2.305

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research,* 2nd edn. New York, NY: The Guilford Press.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull*. 52, 281–302. doi: 10.1037/h0040957

Downey, L., Hayduk, L. A., Curtis, R., and Engelberg, R. A. (2016). Measuring depression-severity in critically ill patients' families with the Patient Health Questionnaire (PHQ): tests for unidimensionality and longitudinal measurement invariance, with implications for CONSORT. *J. Pain Symp. Manag*. 51, 938–946. doi: 10.1016/j.jpainsymman.2015.12.303

Duncan, O. D. (1975). *Introduction to Structural Equation Models*. New York, NY: Academic Press.

Estabrooks, C. A., Huchinson, A. M., Squires, J. E., Birdsell, J., Cummings, G. G., Degner, L., et al. (2009a). Translating research in elder care: an introduction to a study protocol series. *Implement. Sci*. 4:51. doi: 10.1186/1748-5908-4-51

Estabrooks, C. A., Squires, J. E., Cummings, G. G., Birdsell, J., and Norton, P. G. (2009b). Development and assessment of the Alberta Context Tool. *BMC Health Serv. Res.* 9:34. doi: 10.1186/1472-6963-9-234

Estabrooks, C. A., Squires, J. E., Cummings, G. G., Teare, G. F., and Norton, P. G. (2009c). Study protocol for the Translating Research in Elder Care (TREC): building context – an organizational monitoring program in long-term care project (project one). *Implement. Sci.* 4:52. doi: 10.1186/1748-5908-4-52

Estabrooks, C. A., Squires, J. E., Hayduk, L. A., Cummings, G. G., and Norton, P. G. (2011). Advancing the argument for validity of the Alberta Context Tool with healthcare aides in residential long-term care. *BMC Med. Res. Methodol.* 11:107. doi: 10.1186/1471-2288-11-107

Estabrooks, C. A., Squires, J. E., Hayduk, L. A., Morgan, D., Cummings, G. G., Ginsburg, L., et al. (2015). The influence of organizational context on best practice use by care aides in residential long-term care settings. *J. Am. Med. Direct. Assoc.* 16, 537e1–537e10. doi: 10.1016/j.jamda.2015.03.009

Forre, P., and Mooij, J. M. (2019). Causal calculus in the presence of cycles, latent confounders and selection bias. arXiv [Preprint]. *arXiv:1901.00433v1 [stat.ML]*. Available online at: https://arxiv.org/abs/1901.00433 (accessed January 2, 2019

Hardin, A. (2017). A call for theory to support the use of causal-formative indicators: A commentary on Bollen and Diamantopoulos (2017). *Psychol. Methods*. 23, 597–604. doi: 10.1037/met0000115

Hayduk, L. A. (1987). *Structural Equation Modeling With LISREL: Essentials and Advances*. Baltimore: Johns Hopkins University Press.

Hayduk, L. A. (1996). *LISREL Issues, Debates, and Strategies*. Baltimore: Johns Hopkins University Press.

Hayduk, L. A. (2006). Blocked-error R^2: a conceptually improved definition of the proportion of explained variance in models containing loops or correlated residuals. *Quality Quant*. 40, 629–649. doi: 10.1007/s11135-005-1095-4

Hayduk, L. A. (2014a). Seeing perfectly fitting factor models that are causally misspecified: understanding that close-fitting models can be worse. *Edu. Psychol. Measur*. 74, 905–926. doi: 10.1177/0013164414527449

Hayduk, L. A. (2014b). Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing. *BMC Med. Res. Methodol*. 14:124. doi: 10.1186/1471-2288-14-124

Hayduk, L. A., and Glaser, D. N. (2000a) Jiving the four-step, waltzing around factor analysis, and other serious fun. *Struct. Equ. Model.* 7, 1–35. doi: 10.1207/S15328007SEM0701_01

Hayduk, L. A., and Glaser, D. N. (2000b) Doing the four-step, right-2-3, wrong-2-3: a brief reply to Mulaik and Millsap; Bollen; Bentler; and Herting and Costner. *Struct. Equ. Model.* 7, 111–123. doi: 10.1207/S15328007SEM0701_06

Hayduk, L. A., and Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models. *BMC Med. Res. Methodol*. 12:159. doi: 10.1186/1471-2288-12-159

Hayduk, L. A., Pazderka-Robinson, H., Cummings, G. G., Boadu, K., Verbeek, E. L., and Perks, T. A. (2007). The weird world and equally weird measurement models: reactive indicators and the validity revolution. *Struct. Equ. Model.* 14, 280–310. doi: 10.1080/10705510709336747

Heise, D. R. (1975). *Causal Analysis*. New York, NY: John Wiley and Sons.

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., et al. (2014). Common beliefs and reality about PLS: comments on Ronkko and Evermann (2013). *Organizat. Res. Methods* 17, 182–208. doi: 10.1177/1094428114526928

Henseler, J., Ringle, C. M., and Sarstedt, M. (2016). Testing measurement invariance of composites using partial least squares. *Int. Market. Rev*. 33, 405–431. doi: 10.1108/IMR-09-2014-0304

Hubley, A., and Zumbo, B. D. (1996). A dialectic on validity: where we have been and where we are going. *J. General Psychol*. 123, 207–215. doi: 10.1080/00221309.1996.9921273

Joreskog, K. G., and Sorbom, D. (2016). *LISREL 9.1*. Skokie, IL: Scientific Software International Inc.

Little, T. D., Rhemtulla, M., Gibson, K., and Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be one. *Psychol. Methods* 18, 285–300. doi: 10.1037/a0033266

Marsh, H. W., Ludtke, O., Nagengast, B., Morin, A. J. S., and VonDavier, M. (2013). Why item parcels are (almost) never appropriate: two wrongs do not make a right - camouflaging misspecification with item parcels in CFA models. *Psychol. Methods*. 18, 257–284. doi: 10.1037/a0032773

McIntosh, C. N., Edwards, J. R., and Antonakis, J. (2014). Reflections on partial least squares path modeling. *Organizat. Res. Methods* 17, 210–251. doi: 10.1177/1094428114529165

Mulaik, S. A. (2010). *Foundations of Factor Analysis*, 2nd edn. Boca Raton: CRC Press (also New York: Chapman and Hall).

Nagase, M., and Kano, Y. (2017). Identifiability of nonrecursive structural equation models. *Stati. Probabi. Lett.* 122, 109–117. doi: 10.1016/j.spl.2016.11.010

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl, J., and Mackenzie, D. (2018). *The book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Raykov, T., Gabler, S., and Dimitrov, D. M. (2016). Maximal criterion validity and scale criterion validity: a latent variable modeling approach for examining their difference. *Struct. Equ. Model*. 23, 544–554. doi: 10.1080/10705511.2016.1155414

Raykov, T., and Marcoulides, G. A. (2016a). On examining specificity in latent construct indicators. *Struct. Equ. Model*. 23, 845–855. doi: 10.1080/10705511.2016.1175947

Raykov, T., and Marcoulides, G. A. (2016b). Scale reliability evaluation under multiple assumption violations. *Struct. Equ. Model*. 23, 302–313. doi: 10.1080/10705511.2014.938597

Raykov, T., Marcoulides, G. A., Gabler, S., and Lee, Y. (2017). Testing criterion correlations with scale component measurement errors using latent variable modeling. *Structural Equation Modeling*, 24, 468–474. doi: 10.1080/10705511.2015.1136220

Schuberth, F., Henseler, J., and Dijkstra, T. K. (2018). Confirmatory composite analysis. *Front. Psychol.* 9:2541. doi: 10.3389/fpsyg.2018.02541

Song, X.-Y., Lu, Z.-H., Cai, J.-H., and Ip, E. H.-S. (2013). A Bayesian modeling approach for generalizaed semiparametric structural equation models. *Psychometrika* 78, 624–647. doi: 10.1007/s11336-013-9323-7

Wang, Y., Luo, Y., Wang, M., and Miao, H. (2018). Time-invariant biological networks with feedback loops: structural equation models and structural identifiability. *IET Syst. Biol.* 12, 264–272. doi: 10.1049/iet-syb.2018.5004

Zumbo, B. D., and Chan, E. K. H., eds. (2014). *Validity and Validation in Social, Behavioral, and Health Sciences.* Cham: Springer.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 June 2019
doi: 10.3389/fpsyg.2019.01306

Development of a New Instrument for Depression With Cognitive Diagnosis Models

Daxun Wang, Xuliang Gao, Yan Cai* and Dongbo Tu**

School of Psychology, Jiangxi Normal University, Nanchang, China

Most existing instruments for depression are developed based on classical test theory, factor analysis, or sometimes, item response theory, and focus on the accurate measurement of the severity of depressive disorder. Nevertheless, they tend to be less useful in supporting the decision based on *ICD-10* or *DSM-5* because of the lack of detailed information for symptoms. To gain rich and valid information at the symptom level, this article developed a depression test under the framework of cognitive diagnosis models (CDMs), referred to as CDMs-D. A total of 1,181 individuals were finally recruited and their responses were used to examine the psychometric properties of CDMs-D. After excluding poor items for statistical reasons (e.g., low discrimination, poor model-fit or having DIF), 56 items were included in the CDMs-D. The CDMs-D measures all ten symptom criteria for depression defined in *ICD-10* and covers five domains of depression defined by Gibbons et al. (2012). Comparing with the existing self-report measures (such as PHQ-9, SDS, CES-D and so on), a distinguishing feature of the CDMs-D is that it can provide both overall information about the severity of depressive disorder and the assessment information about specific symptoms, which could be useful for diagnostic and interventional purposes.

Keywords: psychological measurement, cognitive diagnosis models, symptom criteria-level information, psychometrics, questionnaires, depression

OPEN ACCESS

Edited by:
Elisa Pedroli,
Istituto Auxologico Italiano (IRCCS),
Italy

Reviewed by:
Klaas Wardenaar,
University Medical Center Groningen,
Netherlands
Roger Muñoz Navarro,
University of Valencia, Spain

****Correspondence:***
Xuliang Gao
gaoxuliang881@qq.com
Yan Cai
cy19791233@aliyun.com
Dongbo Tu
tudongbo@aliyun.com

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *18 February 2019*
Accepted: *20 May 2019*
Published: *04 June 2019*

Citation:
Wang D, Gao X, Cai Y and Tu D
(2019) Development of a New
Instrument for Depression With
Cognitive Diagnosis Models.
Front. Psychol. 10:1306.
doi: 10.3389/fpsyg.2019.01306

INTRODUCTION

Depression is one of the most common and prevalent psychological and behavioral disorders. By the year 2020, depression accounting for 5.7% of the total burden of the disease (Dennis et al., 2016) will be the second disease leading to disability and death with the exception of coronary heart disease according to the World Health Organization (Dennis and Hodnett, 2014). A number of self-report inventories have been developed to assess the severity of the depressive disorder, such as the Self-Rating Depression Scale (SDS; Zung, 1965), the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977) and the Beck Depression Inventory (BDI; Beck et al., 1961).

Despite having sound psychometric properties and being widely used, they are also some rooms for improvement. For example, most existing self-report inventories are unidimensional and yield overall scores indicating the severity of the depressive disorder on a continuum.

To determine whether it is a mild, moderate or severe depression, the scores are compared with some cutoffs. This procedure is straightforward, but it is not informative given that they cannot provide all symptom-level information of depression defined in the 10th revision of the *International Classification of Diseases* (*ICD-10;* World Health Organization [WHO], 2010) or the 5th edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; American Psychiatric Association [APA], 2013). However, these symptom-level information of depression are helpful for assessment, screening, monitoring and even intervention of depression. For example, as shown in **Table 1**, the *ICD-10* groups the symptoms of depression into two sets: typical symptoms and common symptoms and its diagnostic thresholds are specified in terms of the number of symptoms required from each of the two sets. More specially, for the mild depressive episode, two typical symptoms and two common symptoms are required; for the moderate depressive episode, two typical symptoms and at least three common symptoms are required; for the severe depressive episode, all three typical symptoms are present and at least four common symptoms of severe intensity are required. As known, this type assess for depression is more informative than the score cutoffs of conventional inventories given that the patients with the same score may have very different symptoms which can provide more information for screening or treatment.

Form a very different perspective, this study aims to develop a new measure of depression that is aligned with the *ICD-10* to provide more information for the screening and monitoring of depression under the framework of cognitive diagnosis models (CDMs; see Rupp et al., 2010). Compared with the factor analysis technique or item response theory (IRT), the CDMs provide an alternative psychometric framework for test development, psychometric analyses, and score reporting. Although most of research on CDMs lies in the field of education measurement, researchers have been recently aware of their usefulness in psychological disorder assess for identifying individuals' disorder or symptom profiles (e.g., Jaeger et al., 2006; Templin and Henson, 2006; de la Torre et al., 2017). Specifically, it is possible to infer about whether each of the symptom criteria has been satisfied or not from patients' responses to items in an instrument. This information can be useful for screening (or intervening) depressive disorder or other psychological disorders based on the *ICD-10* or *DSM-5*. In addition, compared with factor analysis, CDMs allow latent variables (i.e., symptom criteria) to interact when producing manifest item responses and thus are more flexible.

TABLE 1 | Symptom criteria for depression defined in the *DSM-5* and *ICD-10.*

DSM-5	***ICD-10***
(1) Depressed mood	**Typical symptom criteria**
(2) Markedly diminished interest or pleasure	(1) Depressed mood
(3) Significant weight loss	(2) Loss of interest and enjoyment
(4) Insomnia or hypersomnia	(3) Increased fatigability
(5) Psychomotor agitation or retardation	**Common symptom criteria**
(6) Fatigue or loss of energy	(4) Reduced concentration and attention
(7) Feelings of worthlessness or excessive or inappropriate guilt	(5) Reduced self-esteem and self-confidence
(8) Diminished ability to think or concentrate, or indecisiveness	(6) Ideas of guilt and unworthiness (even in a mild type of episode)
(9) Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan or a suicide attempt or a specific plan for committing suicide.	(7) Bleak and pessimistic views of the future (8) Ideas or acts of self-harm or suicide (9) Disturbed sleep (10) Diminished appetite

In specially, the goal of this study is twofold. First, this study develops a depression test under the framework of CDMs (CDMs-D) based on the *ICD-10* under the CDMs framework, which may be used to assess, screen and monitor depression. Different from the existing self-report questionnaires for depression, the CDMs-D can assess how likely each of the symptom criteria of depression in the *ICD-10* has been met for each patient, and estimate the probability of having mild, moderate and severe depressive episode using the *ICD-10* diagnostic criteria. Second, this study aims to provide an illustration about how CDMs can be used to develop instruments, assess psychometric properties using the *ICD-10* system. This could serve as an example for researchers willing to develop instruments for other psychological disorders using CDMs to provide patient outcomes consistent with *ICD-10* or *DSM-5* criteria.

MATERIALS AND METHODS

Diagnosis System of Depression

Currently, two famous diagnosis systems of depression are *ICD-10* and *DSM-5*, which are both commonly acceptable and used to guide the diagnosis of depression in clinical practice. There are eight common symptom criteria of depressive disorder in *ICD-10* and *DSM-5* (see **Table 1**). In this article, the symptom criteria for depression in the *ICD-10* were used in that the *ICD-10* distinguishes three types of depression (mild, moderate or severe/major depression) and thus could provide more information.

Cognitive Diagnosis Models

In the context of CDMs, 10 symptom criteria of depression in *ICD-10* are treated as latent variables that need to be measured, each with two outcomes – 1 and 0, representing presence and absence, respectively. Based on individuals' responses to items of the CDMs-D and the aforementioned item and symptom association matrix, CDMs estimate the symptom profile for each individual. For example, if the symptom profile for an individual is estimated to be (0,1,1,0,0,0,1,1,0,0), this individual is said to meet symptom criteria 2, 3, 7, and 8. In addition, CDMs can also estimate the probability of an individual meets each criterion.

An array of CDMs can be found in the literature (Rupp et al., 2010). In this study we adopt the generalized deterministic input, noisy, "and" gate (G-DINA; de la Torre, 2011) model framework because (1) it is one of the most general CDMs with many applications and (2) it is very flexible and subsumes many reduced CDMs. The G-DINA model, like most other CDMs, is a psychometric model specifying how individuals respond to each item given their symptom criteria. Take item "I feel worthless and ashamed" as an example, which measures (C5) "reduced self-esteem and self-confidence" and (C6) "ideas of guilt and unworthiness."

Let $\alpha = (\alpha_1, \alpha_2)$ denote the profile of these two criteria. Based on the G-DINA model (de la Torre, 2011), the probability of endorsement on this item given the symptom profile α can be written by $P(\alpha) = \phi_0 + \phi_1\alpha_1 + \phi_2\alpha_2 + \phi_{12}\alpha_1\alpha_2$. More specifically, for α = (0,0), where both symptoms are absent, the corresponding endorsement probability is$P(0, 0) = \phi_0$; for α = (1,0), where symptom C5 is present but C6 is absent, the corresponding endorsement probability is $P(1, 0) = \phi_0 + \phi_1$, where ϕ_1 is the effect of symptom C5; for α = (0,1), where symptom C5 is absent but C6 is present, the corresponding endorsement probability is $P(0, 1) = \phi_0 + \phi_2$, where ϕ_2 is the effect of symptom C6; and for α = (1,1), where both symptoms are present, the corresponding endorsement probability is $P(1, 1) = \phi_0 + \phi_1 + \phi_2 + \phi_{12}$, where ϕ_{12} is the interaction effect of symptoms C5 and C6.

Although the G-DINA model considers all possible interactions among measured symptom criteria, researchers may have some assumptions about how symptom criteria produce item responses. For example, the deterministic inputs, noisy "and" gate (DINA) model assumes that the endorsement probability will not increase unless all measured symptom criteria have been present. This model can be obtained, for the aforementioned example, by setting $\phi_1 = \phi_2 = 0$ such that $P(0, 0) = P(1, 0) = P(0, 1) = \phi_0$ and $P(1, 1) = \phi_0 + \phi_{12}$. In contrast, the deterministic inputs, noisy "or" gate (DINO; Templin and Henson, 2006) model assumes that a high endorsement probability is expected if any of the measured symptom criteria is present. This model can be obtained by setting $\phi_1 = \phi_2 = -\phi_{12}$ such that $P(0, 0) = \phi_0$ and $P(1, 0) = P(0, 1) = P(1, 1) = \phi_0 + \phi_1$. In addition, the addictive CDM (A-CDM; de la Torre, 2011), linear logistic model (LLM; Maris, 1999) and reduced reparameterized unified model (rRUM; Hartz et al., 2002) can be obtained by assuming all symptom criteria contribute independently and uniquely without interaction effects. For more details on these models, please refer to de la Torre (2011).

Development of Cognitive Diagnostic Test for Depression (CDMs-D)

The CDMs-D is designed to be a self-report instrument and the ultimate goal is to infer whether an individual has satisfied each of the symptom criteria of depression defined in the *ICD-10* and the probability of having mild, moderate and severe depressive episode from his or her responses. The CDMs-D primitively included 89 items which were carefully chosen according to the depression symptom criteria in the *ICD-10* from several self-rating inventories, including the Zung's SDS, the CES-D (Radloff, 1977), the Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001), the Hospital Anxiety Depression Scale (HADS), Carroll's Depression Scale (CDS; Carroll et al., 1981), Minnesota Multiphasic Personality Inventory (MMPI; Hathaway and McKinley, 1942), the Brief Depression Scale (BDS; Koenig et al., 1992), the Geriatric Depression Scale (GDS), the Edinburgh postnatal depression Scale (EPDS; Cox et al., 1987) and the Adolescents Depression Emotion Self-assessment Scale (ADESC; Huang et al., 2004). The chosen 89 items measure all ten depression symptom criteria in *ICD-10* and involve five domains of depression defined by Gibbons et al. (2012), namely, mood (14 items), cognition (30 items), behavior (21 items), somatic complaints (17 items) and ideas or acts of suicidality (7 items). Items were revised to refer to the previous 2-week period and to have consistent response categories. Each item measures at least one depression symptom criterion in *ICD-10*.

The way of an individual responding to an item can be reasonably assumed to be influenced by whether she/he has satisfied some symptom criteria. For example, an individual may agree with that "I feel worthless and ashamed" if she/he has "reduced self-esteem and self-confidence" (C5) or "ideas of guilt and unworthiness" (C6) and agree with that "I wish to be dead" if she has "ideas or acts of self-harm or suicide" (C8). To make inference as to whether individuals have satisfied each symptom criterion from their item responses, an item by symptom association matrix giving which symptom criteria may influence individuals' item responses needs to be developed in advance. For CDMs-D, the item and symptom association matrix was constructed using the Delphi method with three experts (two psychotherapists with more than 5 years of clinical experience and one with 5-year research experience in the measurement of depression). **Table 2** gives some exemplary items and their association with symptom criteria, where entry 1 indicates a symptom criterion is measured by the item and entry 0 indicates not. On average, each item measures 1.67 symptom criteria, and each criterion is measured by 14.9 items.

Participant Sample

Participants include healthy individuals and patients with depression. Depressive patients, who were being treated for depression, were recruited from eight health centers and hospitals in seven provinces/cities of China, whereas the healthy individuals were mainly from colleges and social groups. The selected seven provinces/cities distribute in east, south, west, and

TABLE 2 | Exemplary items in CDMs-D.

Items (abbreviated content)	Domain	Q-matrix									
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Worthlessness and shame	Cognition	0	0	0	0	1	1	0	0	0	0
Feeling unhappy	Mood	1	0	0	0	0	0	0	0	0	0
Everything is laborious	Behavior	0	0	1	0	0	0	0	0	0	0
Wish to be dead	Suicidality	0	0	0	0	0	0	0	1	0	0

Criteria C1, C2, and C3 represent three typical symptoms; criteria C4–C10 represent seven common criteria in ICD-10.

north area of China and covers mainly area of China. The final selection of both depressive patients and healthy individuals were recruited according to the following exclusion criteria: history of psychosis, schizoaffective disorder, or schizophrenia; organic neuropsychiatric syndrome, such as dementia and Parkinson disease; drug or alcohol dependence over the past 3 months, but not excluded patients with episodic abuse related to mood episodes. The study also had exclusion criteria to screen the healthy individuals: history of psychosis, schizoaffective disorder, or schizophrenia; any diagnosis or treatment for psychiatric illness over the past year. The study was approved by the medical ethics committees of participating health center and hospitals, and all participants were provided written informed consent.

A total of 1,286 samples were recruited, among which 92 samples had large missing data in the questionnaire and 13 samples met the exclusion criteria. After excluding the above 105 samples, the final selected participant sample was consisted of 1,181 individuals aged from 18 to 80 with mean = 31.8 (*SD* = 12.92) based on the above exclusion criteria for this study. The number of depressive patients and healthy individuals were 488 (41.3%) aged from 18 to 80 with mean = 36.8 (*SD* = 14.9), and 693 (58.7%) aged from 18 to 57 with mean = 28.36 (*SD* = 10.03), respectively.

The total sample was randomly split into two subsamples. One of the resulting two subsamples was half of the overall sample and used as a calibration sample (N_1 = 591) to develop the CDMs-D. The other half sample was used as the cross validation sample (N_2 = 590) to verify the CDMs-D and investigate the reliability and validity of CDMs-D. Detailed demographic information was documented in **Table 3**.

TABLE 3 | Demographic characteristics of depressive disorder patients and healthy individuals.

Characteristic	Calibration Sample, % (N_1 = 591) Total (Male/Female)	Validation Sample, % (N_2 = 590) Total (Male/Female)
Gender		
Male	46.4	47.8
Female	53.6	52.2
Age, years old		
18–29	62.1 (46.3/53.7)	58.5 (49.9/50.1)
30–39	13.5 (55/45)	12.2 (38.9/61.1)
40–49	12.5 (46.6/53.4)	12.7 (50.7/49.3)
50–59	8.8 (42.3/57.7)	12.0 (45.1/54.9)
≥60	3.0 (27.8/72.2)	4.6 (34.6/65.4)
Education		
Some high school or < 9th grade	20.6 (43/57)	20.5 (38.3/61.7)
High School diploma or GED	17.3 (38.9/61.1)	19.5 (47/53)
College graduate	52.5 (50.6/49.4)	51.2 (50.2/49.8)
Graduate or professional degree	9.6 (42.1/57.9)	7.1 (52.4/47.6)
Area		
Rural	48.4 (52.3/47.7)	48.1 (53.9/46.1)
Urban	51.4 (46.2/53.8)	50.2 (41.4/58.6)
other	0.2 (0/1)	1.7 (45.4/54.6)
Group		
Healthy	58.2 (47.4/52.6)	59.2 (50.9/49.4)
Depression	41.8 (45.5/54.5)	40.8 (42.7/57.3)

Statistical Analysis

The calibration sample (N_1 = 591) was used in this step to develop the CDMs-D.

Item Analysis

Selecting suitable CDM is deemed to be a critical procedure for making valid inferences. Although a number of CDMs are available, it's not always clear which model should be chosen for a given data set. The Wald test (de la Torre, 2011; Ma et al., 2016) was proposed to evaluate whether the reduced CDM can be replaced by the saturated CDM without significant loss in model-fit (de la Torre, 2011), and the results of Ma et al. (2016) indicated that the chosen CDMs via the Wald test performed better than the saturated CDM in terms of estimation of person parameter. In this study five special or reduced CDMs were considered, which were the deterministic inputs, noisy "and" gate model (DINA; Junker and Sijtsma, 2001), the deterministic input, noisy "or" gate model (DINO; Templin and Henson, 2006), the addictive CDM (A-CDM; de la Torre, 2011), the linear logistic model (LLM; Maris, 1999) and the reduced reparameterized unified model (RRUM; Hartz et al., 2002). The Wald test was carried out for items measuring more than one criterion in that all CDMs are equivalent for single criterion items.

After choosing the suitable model for each item, the *S*-X^2 item fit statistic (Orlando and Thissen, 2000) was used to assess the adequacy of item fit, followed by the detection of the differential item functioning (DIF) for different groups (e.g., female and male, rural and urban) using the Wald statistic (Hou et al., 2014). Then, the discrimination index (*Disc*) suggested by de la Torre (2008) was calculated to assess item quality. The above statistical analyses were conducted step by step.

In Step 1, the item fit analysis was carried out via *S*-X^2 item fit statistic and items with poor fit (*p*-value of *S*-X^2 less than 0.01) were deleted from the CDMs-D. In Step 2, for the remainder items in Step 1, DIF analysis was employed and items with DIF were excluded from the CDMs-D. In Step 3, for the remainder items in Step 2, we assessed item discrimination and items with low discrimination (*Disc* < 0.4) were deleted. That is to say, any item that had low discrimination (*Disc* < 0.4), had DIF or fitted to the data inadequately was removed from the CDMs-D. This procedure (three steps) was repeated until no item was deleted. The GDINA R package (Ma and de la Torre, 2016) and Custom-written code in R (R Core Team, 2016) were used for analyses.

Then the cross validation sample (N_2 = 590) was used to re-analyze and validate the remained items selected by the calibration sample (N_1 = 591). At this step the items that had low discrimination, DIF or poor item fit would be also deleted form the final CDT-T.

Reliability and Validity

The analysis of both the reliability and validity were carried out for the final CTD-D after above item analysis and item selection only with the cross validation sample (N_2 = 590). Under the framework of cognitive diagnosis, the symptom-level classification consistency and accuracy indices (Cui et al., 2012; Templin and Bradshaw, 2013) based on CDMs were investigated for CDMs-D. Criterion-related and convergent validity were then assessed by the coefficients of correlation between the CDMs-D and the SDS and individual's self-reported depression and the. Content validity was examined as well in terms of whether the CDMs-D measures all the depression symptoms defined in *ICD-10* and covers all the domains of depression defined by Gibbons et al. (2012).

Depression Assessment

The posterior probability of satisfying symptom criterion k for individual i can be calculated as in

$$P(\alpha_k|X_i) = \sum_{\forall w:\alpha_{wt}=1} P(\alpha_w|X_i),$$

where $P(\alpha_w|X_i)$ is the posterior probability of having symptom profile α_w for individual$i$. Based on the posterior probability of satisfying each symptom criterion, we can calculate the probability of having each symptom criteria profile and the probability of being considered as mild, moderate or severe depression.

RESULTS

Item Analysis of the CDMs-D

Using the aforementioned item analysis procedure, 31 items were deleted with the calibration sample (N_1 = 591). Specifically, 20 of them had low discrimination index (*Disc* < 0.4), 5 were DIF items and 10 showed poor item-fit (p < 0.01). After that, the remained 58 items were analyzed with the cross validation sample (N_1 = 590). Results showed that 56 items had high discrimination, good item-fit and no DIF except two items with low item fit. Therefore, the final CDMs-D had 56 items, which are given in **Table 4**. The CDMs-D measures all ten symptom criteria for depression defined in the *ICD-10* and involves five domains of depression which are mood (7 items), cognition (23 items), behavior (10 items), somatic complaints (9 items) and ideas or acts of suicidality (7 items). The number of items measuring each symptom criteria varies from 4 to 22 with an average of 10.4. In addition, there are 17, 31, 7, and 1 item (s) measuring 1, 2, 3, and 4 symptom criteria respectively with an average of 1.85 symptom criteria per item.

Reliability and Validity

Classification consistency refers to the extent to which participant classifications agree between two independent administrations, which is also called the reliability of classifications (Cui et al., 2012). As shown in **Table 5**, all attributes have classification consistency greater than 0.95 which suggests the CDMs-D has high reliability of classifications. In addition, classification accuracy refers to the extent to which the participants' classifications agree with their true latent classes (Cui et al., 2012). **Table 5** showed that the CDMs-D had high probability of classifying participants accurately based on their observed responses since all attributes have classification accuracy greater than 0.94.

From **Table 4**, the CDMs-D measures all depression symptoms defined in *ICD-10* and cover all five domains of depression defined by Gibbons et al. (2012), which implies that it has appropriate content validity. As for the criterion-related and convergent validity, the CDMs-D has a correlation of 0.707 (p < 0.001) and 0.810 (p < 0.001) with self-reported depression and SDS, respectively. The estimated probability of having mild, moderate or severe depression has a correlation of 0.791 (p < 0.001) and 0.651 (p < 0.001) with SDS and self-reported depression, respectively. Moreover, we calculated the coefficient of classification consistency between the CDMs-D and the structured clinical interview by psychotherapists via *ICD-10*, and results showed that there had a moderate coefficient of classification consistency with 0.463 (p < 0.001) between them. **Figures 1**, **2** show the 95% confidence intervals (CIs) for the mean CDMs-D score and the mean probability of having depressive disorder, respectively, for individuals with or without depression defined by the SDS or self-reported depression. Different groups have quite different mean CDMs-D scores and mean probabilities of depressive disorder, suggesting that the CDMs-D has the power to discriminate individuals with depression at different levels of severity.

Screening Scores Reporting

Compared with existing instruments for depression, CDMs-D could provide unique screening information for each patient. For illustration, score reports for four individuals (three patients and one healthy individual) were displayed in **Figure 3**. Three patients were chosen in that: (1) they were classified as moderate depression by their psychotherapists; (2) they had the same SDS score and were defined as moderate depression via the criterion of SDS; (3) they reported that they usually had considerable difficulty in continuing with social, work or domestic activities. **Figure 3** shows the posterior probability that each criterion has been satisfied for these individuals. Based on these probabilities, the chances of having mild, moderate or severe depression for each individual can be calculated.

Individual A (male, 25 years old and from rural) has very high posterior probabilities of satisfying the typical symptom C2 and the common symptoms C10. Based on *ICD-10*, the estimated probabilities of being normal, mild, moderate and severe depression are 0.81, 0.12, 0.06, and 0.01, respectively, which suggests that it is unlikely for him to have depressive disorder.

Patients B, C, and D are all classified as having moderate depressive disorder by the CDMs-D (with the estimated posterior probability of 0.99, 0.99, 0.63, respectively), which is consistent to the results of their psychotherapists and SDS. However, they differ in their symptom profiles. From **Figure 3**, Patient B (female, 23 years old and from rural) probably satisfies two typical symptoms (C1 and C3) and four common symptoms (C4, C5, C7, and C8); Patient C (male, 29 years old and from

TABLE 4 | Final items of the CDMs-D.

Item No.	Item abbreviation	Selected model	Discr.	Item-fit		DIF1		DIF2		Domain of depression
				$S\text{-}X^2$	*p*	*Wald*	*p*	*Wald*	*p*	
Item 1	Lose daily life ability	RRUM	0.57	55.89	0.232	0.90	0.827	1.24	0.743	Behavior
Item 2	Talking slow and dully	DINA	0.54	62.32	0.096	3.07	0.216	0.39	0.822	Behavior
Item 7	Unable to do things	DINA	0.58	56.06	0.227	1.68	0.431	0.48	0.788	Behavior
Item 8	Broken down	ACDM	0.94	40.41	0.774	0.26	0.992	2.78	0.595	Behavior
Item 9	Full of energy	ACDM	0.60	42.02	0.750	3.08	0.380	4.32	0.229	Behavior
Item 10	Talk less	ACDM	0.58	38.38	0.863	2.26	0.520	1.13	0.770	Behavior
Item 11	Confidence of doing everything	ACDM	0.51	50.73	0.366	2.89	0.576	1.67	0.796	Cognition
Item 12	Loneliness feelings	GDINA	0.44	57.62	0.214	0.45	0.797	4.25	0.120	Cognition
Item 13	Clear mind	GDINA	0.62	37.42	0.906	0.75	0.687	4.04	0.133	Cognition
Item 14	Unchanged working/studying ability	ACDM	0.60	51.05	0.393	1.08	0.781	2.54	0.468	Cognition
Item 15	Deal with daily life easily	RRUM	0.54	65.04	0.062	1.53	0.675	5.11	0.164	Cognition
Item 16	Future hopeless feelings	ACDM	0.82	46.62	0.530	2.89	0.576	0.56	0.967	Cognition
Item 17	Unpopularity feelings	GDINA	0.56	40.00	0.843	1.27	0.530	0.43	0.806	Cognition
Item 18	Future hopeful feelings	RRUM	0.76	51.25	0.386	1.75	0.625	3.27	0.351	Cognition
Item 19	Loser feelings	ACDM	0.87	42.40	0.736	2.89	0.408	5.40	0.145	Cognition
Item 20	Failure in life	ACDM	0.84	39.60	0.829	0.70	0.873	1.53	0.676	Cognition
Item 21	Worthiness	ACDM	0.63	48.95	0.475	0.26	0.968	1.60	0.660	Cognition
Item 22	Worthlessness and shame	ACDM	0.89	40.20	0.810	0.96	0.811	0.01	1.000	Cognition
Item 23	Guilty feelings	GDINA	0.57	62.77	0.106	0.16	0.924	0.94	0.626	Cognition
Item 24	Reading disorder	ACDM	0.58	57.90	0.180	4.82	0.186	3.61	0.307	Cognition
Item 25	Feelings of being talking about	DINO	0.40	59.59	0.143	0.25	0.882	2.15	0.342	Cognition
Item 26	Concentration difficulty	ACDM	0.52	37.11	0.894	0.99	0.803	2.43	0.488	Cognition
Item 27	Pessimism about future	ACDM	0.91	41.29	0.743	5.90	0.207	1.33	0.856	Cognition
Item 28	Clear and quick thinking	GDINA	0.73	46.00	0.635	2.10	0.349	0.30	0.860	Cognition
Item 29	Judicious	GDINA	0.46	46.25	0.625	0.17	0.917	2.00	0.367	Cognition
Item 30	Desperation	ACDM	0.88	32.76	0.964	3.42	0.331	1.55	0.671	Cognition
Item 31	Disappointing	ACDM	0.76	66.10	0.052	2.01	0.570	3.10	0.376	Cognition
Item 32	Efforts are useless	RRUM	0.75	55.75	0.236	4.04	0.257	7.01	0.072	Cognition
Item 33	Past sorrow	ACDM	0.55	45.13	0.631	0.88	0.831	4.93	0.177	Cognition
Item 34	Sex with joy	ACDM	0.44	72.58	0.016	2.27	0.518	0.51	0.917	Mood
Item 35	Abandonment	ACDM	0.74	48.52	0.452	3.32	0.506	0.36	0.986	Mood
Item 36	Still depressed with others' help	ACDM	0.75	54.02	0.288	0.59	0.898	1.69	0.640	Mood
Item 37	Inner mental collapsion	GDINA	0.81	43.58	0.654	3.34	0.912	2.97	0.936	Mood
Item 38	Satisfaction feelings	ACDM	0.74	49.40	0.457	3.34	0.342	0.37	0.946	Mood
Item 39	Fond of communication	RRUM	0.59	52.29	0.348	1.98	0.577	1.53	0.675	Mood
Item 40	Loss of interest	ACDM	0.72	69.46	0.029	0.37	0.946	2.56	0.464	Mood
Item 41	Early awakening	GDINA	0.52	43.90	0.715	0.77	0.680	4.14	0.126	Somatic
Item 42	Loss of appetite	GDINA	0.48	62.27	0.114	0.30	0.862	0.58	0.748	Somatic
Item 43	Awakening at nights	GDINA	0.51	57.70	0.212	0.02	0.990	1.40	0.496	Somatic
Item 44	Poor quality of sleep	GDINA	0.66	54.13	0.320	0.65	0.721	1.77	0.412	Somatic
Item 45	Poor sleep or somnolence	GDINA	0.59	49.58	0.490	1.43	0.489	2.07	0.356	Somatic
Item 46	Dizziness	GDINA	0.62	58.33	0.196	1.30	0.522	0.21	0.900	Somatic
Item 47	Good appetite	GDINA	0.93	74.64	0.014	1.81	0.404	0.92	0.632	Somatic
Item 48	Unchanged appetite	RRUM	0.82	71.69	0.019	3.72	0.293	0.72	0.869	Somatic
Item 49	Insomnia-early	GDINA	0.54	57.35	0.221	1.58	0.453	0.22	0.897	Somatic
Item 50	Suicidal thoughts	ACDM	0.83	41.60	0.765	3.94	0.268	0.90	0.827	Suicidality
Item 51	Inability to continue	ACDM	0.86	50.01	0.433	0.53	0.911	1.02	0.796	Suicidality
Item 52	Hardship feelings	ACDM	0.77	42.26	0.741	6.69	0.083	2.46	0.483	Suicidality
Item 53	Planning suicide	GDINA	0.64	37.90	0.895	0.42	0.811	1.50	0.473	Suicidality
Item 54	Wish was dead	GDINA	0.81	36.23	0.928	0.11	0.947	0.94	0.626	Suicidality
Item 55	Life is meaningful	ACDM	0.88	60.72	0.086	2.31	0.805	5.38	0.371	Suicidality
Item 56	Others' life will be better without me	ACDM	0.69	39.02	0.845	1.09	0.780	2.42	0.490	Suicidality

Disc., discrimination; DIF1, DIF (Male vs. Female); DIF2, DIF(Rural vs. Urban); RRUM, the Reduced Reparameterized Unified Model; G-DINA, the general DINA model; A-CDM, the additive CDM.

TABLE 5 | The reliability and validity of the CDMs-D.

CDMs-D	Classification consistency	Classification accuracy	Test score of CDMs-D	SDS	Self-reported depression
Test score of CDMs-D	—	—	1	0.810***	0.707***
Screening assessments	—	—	0.902***	0.791***	0.651***
C1	0.983	0.973	0.827***	0.689***	0.621***
C2	0.961	0.978	0.669***	0.692***	0.516***
C3	0.973	0.960	0.805***	0.699***	0.590***
C4	0.958	0.938	0.705***	0.722***	0.470***
C5	0.962	0.958	0.763***	0.603***	0.486***
C6	0.970	0.955	0.776***	0.647***	0.503***
C7	0.962	0.959	0.765***	0.742***	0.552***
C8	0.987	0.973	0.705***	0.621***	0.539***
C9	0.965	0.945	0.624***	0.563***	0.464***
C10	0.977	0.952	0.583***	0.608***	0.445***

SDS, Zung's Self-Rating Depression Scale (1965); CDS, Carroll's Depression Scale (Carroll et al., 1981); CES-D, Center for Epidemiologic Studies Depression Scale (Radloff, 1977); C1–C10 represent 10 symptom criteria for depression defined in ICD-10 shown in ***Table 1****; ***represents p < 0.001. screening assessments were the probability of depressive disorder based on CDMs-D via CDM.*

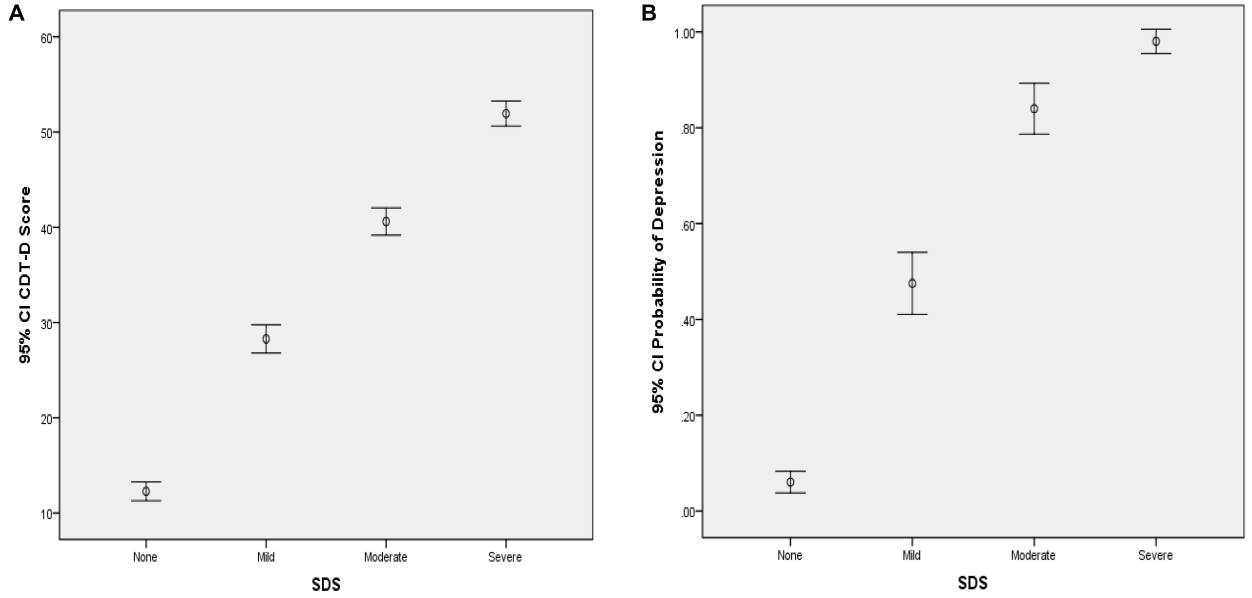


FIGURE 1 | Error bar graph of the CDMs-D scores **(A)** and the probability of depressive disorder **(B)** for different groups via SDS. 95% CI, 95% confidence interval. The probability of depressive disorder (i.e., probability of mild, moderate and severe depression) was calculated based on the CDMs-D and the diagnostic criteria in *ICD-10* via CDMs.

rural) probably satisfies two typical symptoms (C1 and C3) and four common symptoms (C5, C6, C9, and C10); and Patient D (male, 58 years old and from urban) probably satisfies two typical symptoms (C1 and C3) and five common symptoms (C4, C5, C6, C7, and C9). Additionally, it can be seen that Patient B has a very high posterior probability of having symptom C8 (ideas or acts of self-harm or suicide) but Patient C and Patient D have very low probabilities. The information of symptom spectrum of each individual as showed in **Figure 3** give insight into tailoring individual-specific treatments for depression. For example, for Patient B, the targeted treatment should focus on decreasing the chance of having ideas or acts of self-harm or suicide, for Patient C the targeted treatment should aim to decrease the fatigability and improve the enjoyment, while for Patient D, helping her to establish a brief of bring future is very important for him.

DISCUSSION AND CONCLUSION

In this article, a new instrument for depression, the CDMs-D, is developed under the CDM framework based on *ICD-10*. This is the first study to measure the depressive disorder from the CDM perspective, though CDMs have been used as psychometric tools to analyze patient-reported outcomes, such as the pathological gambler in Templin and Henson (2006),

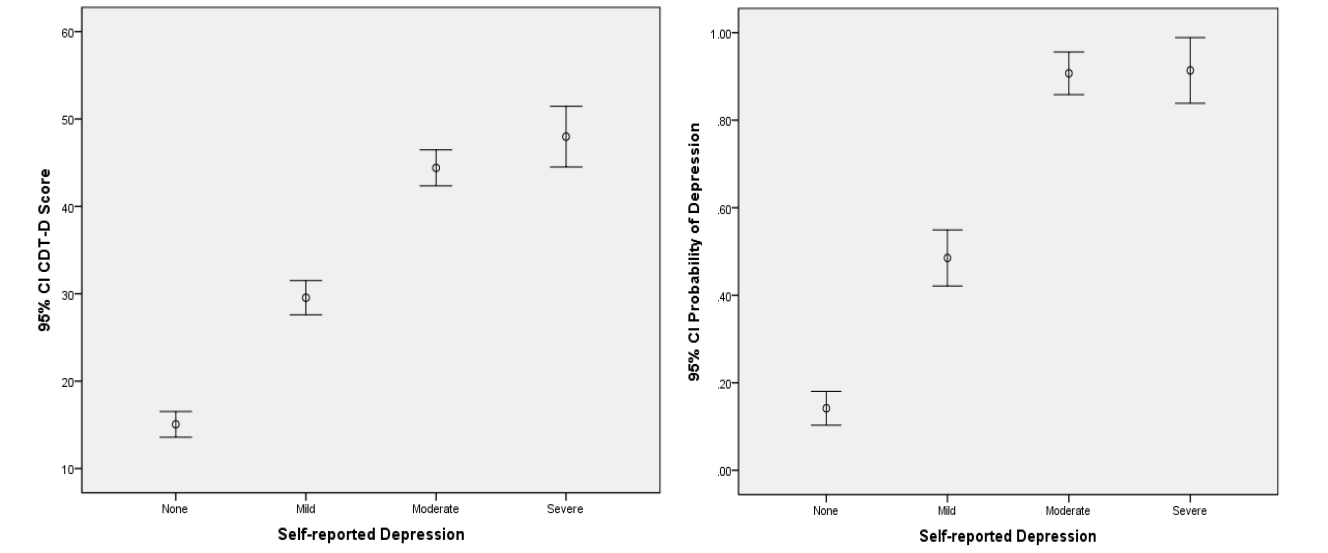


FIGURE 2 | Error bar graph of the CDMs-D scores and the probability of depressive disorder for different groups via self-reported depression. 95% CI, 95% confidence interval. The probability of depressive disorder (i.e., probability of mild, moderate, and severe depression) was calculated based on the CDMs-D and the diagnostic criteria in *ICD-10* via CDMs.

neurocognitive functions in schizophrenia in Jaeger et al. (2006), internet addition in Tu et al. (2017) and the Millon Clinical Multiaxial Inventory-III in de la Torre et al. (2017). CDMs provide a set of psychometric tools to assess item properties, test reliability (Cui et al., 2012) and validity, and in this study, the CDMs-D with 56 items has been shown to have good reliability and validity. Comparing with the existing self-report measures (such as SDS, CES-D), one outstanding advantage of the new measure is that it measures all symptom criteria defined in the *ICD-10* and can provide symptom level reports. In addition, the high correlation between the CDMs-D and SDS indicated that the general-level information of depression they provided were high consistent. However the CDMs-D can provide the additional symptom-level information of depression. This dues to that the CDMs have the unique feature that can provide rich information in terms of whether the participants have met each symptom and of estimating the probability of having mild, moderate, and severe depressive disorder. Such information tends to be superior to the decision made based on total scores from some existing questionnaire in that it is obtained according to the ICD-10.

The proposed measure also has some latent contributions for the specifically assessing/screening for *ICD* and *DSM*-based depression. For example, this proposed measure aims to screen and monitor ICD and DSM-based depression, therefore it may provide a beneficial supplement to a clinician, especially when the patients cannot clearly and directly report whether all the symptoms defined in DSM or ICD are present. Another latent contribution is that it may reduce the burden of a clinician when there are large subjects for screening or monitoring. Moreover, a patient can conveniently make a self-examination about *ICD* and *DSM*-based depression by using the CDMs-D. Finally, a clinician can use the information from the measure, the clinical interview and others together to make diagnosis.

It is the CDMs that make these inferences possible, but the CDMs need to be used with cautions. Unlike classical test theory, factor analysis and IRT models, CDMs typically assume that latent variables are binary (Rupp et al., 2010). Because of this assumption, CDMs lend themselves well to modeling symptoms for many disorders in psychiatry. However, it is reasonable to ask whether the symptoms are binary or not in nature. It should be noted that all psychometric models, including CDMs, are just approximations of the real world, and therefore, as long as the symptoms can be approximately treated as binary variables especially for the ICD and DSM-based assessment of depression, the inferences can be useful. Additionally, CDMs consider the complex interactions among latent binary variables (de la Torre, 2011; Templin and Bradshaw, 2014) (e.g., unobserved symptoms). This, on one hand, allows greater flexibility than most IRT models in modeling item responses; but, on the other hand, tends to make the model complex with, sometimes, too many parameters. This study considered simplifying the saturated CDM with all possible interactions to some reduced models with fewer parameters to obtain more stable parameter estimates. These analyses are important because, in general, a simpler model should be preferred to a complicated model if both fit data well.

Despite promising results, to unlock the potential of the CDMs, more research is needed. First, the current CDMs-D with 56 items is relatively long. It is important to consider a shorter version of CDMs-D to decrease patients' test burden (Smiits et al., 2011). The computerized adaptive testing (CAT) may be an option to decrease the test length without a loss of measurement precision. Some research on combining

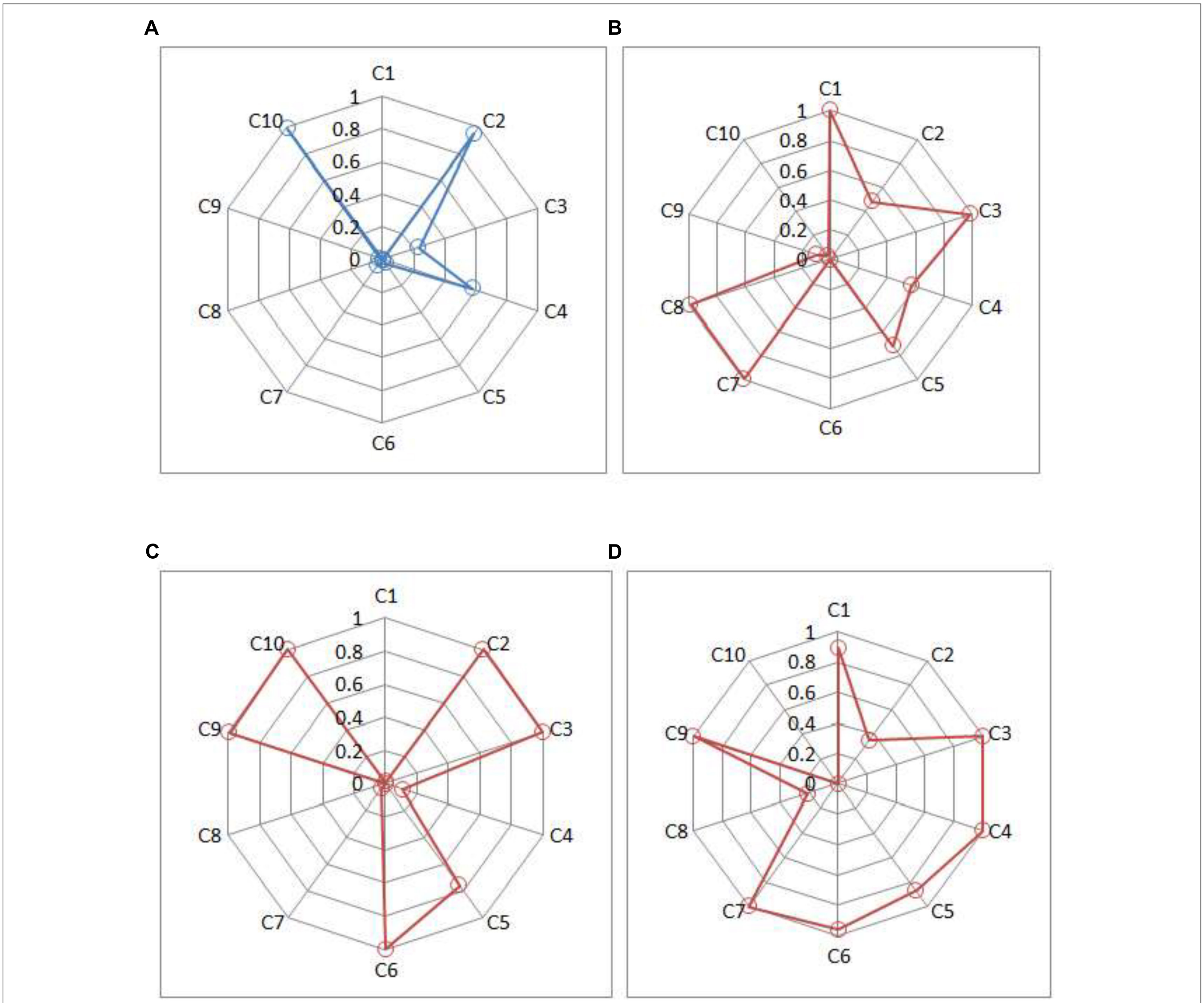


FIGURE 3 | Symptom spectrum of depression for three patients and one healthy individual. **(A)** Individual A, **(B)** Patient B, **(C)** Patient C, and **(D)** Patient D. Criteria C1, C2, and C3 represent three typical symptoms; criteria C4–C10 represent seven common criteria in *ICD-10* in **Table 1**.

CDM and CAT can be found in literature in the field of psychometrics (e.g., Cheng, 2009), but applications are lagging behind. Therefore, further research may empirical investigate how to amalgamate CDMs and CAT (CD-CAT; Cheng, 2009; Wang et al., 2011) to develop the CAT version of CDMs-D. Second, the outputs with probabilities of the proposed measure may be not familiar and accustomed for users. For example, this CDT-T may provide two types of probabilities: one is the probabilities of none depression, mild depression, moderate depression and severe depression, which add up to 100%; another is the probability of presence for each symptom. The former probabilities can be used as screening or monitoring while the latter probabilities can be used to investigate the symptoms characteristic for each patient. That is to say this measure can provide both general level and symptom level information. Third, this article considered the symptom criteria for depression defined in *ICD-10*, future research may explore whether it is appropriate to use the criteria defined in *DSM-5*. Fourth, future study should compare the CDMs-D and the structured interview protocols based on either the *ICD-10* or the *DSM-5*. Fifth, except of results in CDMs-D, other evidences such as a structured clinical interview should also be taken full consideration to give a diagnosis of depression. Sixth, there are also some commonly used dimensional measures of depression that are not included in this article, therefore more measures should be considered for future study. Last, the selected CDMs in this study involve a large number of parameters. The sample used for test calibration may not be large enough and therefore, some statistical procedures such as the Wald test for model selection and DIF detection may be affected due to poorly estimated covariance matrix (Philipp et al., 2017). Larger sample should be considered to stabilize the parameter estimation.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of ethics committee of Center for Mental Health Education and Research of Jiangxi Normal University with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the ethics committee of Center for Mental Health Education and Research of Jiangxi Normal University.

AUTHOR CONTRIBUTIONS

DW contributed to thesis writing and code writing. XG processed the data. YC performed to guide the data processing and code writing. DT contributed to guide the thesis writing and code writing.

FUNDING

This work was supported by the National Natural Science Foundation of China (31660278 and 31760288), and the graduate student innovation fund of Jiangxi Normal University (YC2018-B025).

REFERENCES

American Psychiatric Association [APA] (2013). *Diagnostic and Statistical Manual of Mental Disorders*. 5th Edn. Arlington, VA: American Psychiatric Publishing.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., and Erbaugh, J. (1961). An inventory for measuring depression. *Arch. Gen. Psychiatry* 4, 561–571. doi: 10.1001/archpsyc.1961.01710120031004

Carroll, B. J., Feinberg, M., Smouse, P. E., Rawson, S. G., and Greden, J. F. (1981). The Carroll rating scale for depression. I. Development, reliability and validation. *Br. J. Psychiatry J. Men. Sci.* 138, 194–200. doi: 10.1192/bjp.138.3.194

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika* 74, 619–632. doi: 10.1007/s11336-009-9123-2

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R foundation for Statistical Computing.

Cox, J. L., Holden, J. M., and Sagovsky, R. (1987). Detection of postnatal depression. development of the 10-item Edinburgh postnatal depression scale. *Br. J. Psychiatry J. Men. Sci.* 150, 782. doi: 10.1192/bjp.150.6.782

Cui, Y., Gierl, M. J., and Chang, H. H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *J. Educ. Meas.* 49, 19–38. doi: 10.1111/j.1745-3984.2011.00158.x

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: development and applications. *J. Educ. Meas.* 45, 343–362. doi: 10.1111/j.1745-3984.2008.00069.x

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika* 76, 179–199. doi: 10.1007/s11336-011-9207-7

de la Torre, J., van der Ark, A., and Rossi, G. (2017). Analysis of clinical data form cognitive diagnosis modeling framework. *Meas. Eval. Counsel. Dev.* 1, 1–16.

Dennis, C. L., Brown, H. K., and Morrell, J. (2016). *Interventions (Other Than Psychosocial, Psychological and Pharmacological) for Preventing Postpartum Depression*. Hoboken, NJ: John Wiley & Sons, Ltd.

Dennis, C. L., and Hodnett, E. (2014). Psychosocial and psychological interventions for treating postpartum depression. *Cochrane Database Syst. Rev.* 89:92.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., et al. (2012). Development of a computerized adaptive test for depression. *Arch. Gen. Psychiatry* 69, 1104–1112.

Hartz, S., Roussos, L., and Stout, W. (2002). *A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality*. doctoral dissertation. Champaign: University of Illinois at Urbana.

Hathaway, S. R., and McKinley, J. C. (1942). A multiphasic personality schedule (Minnesota): III. The measurement of symptomatic depression. *J. Psychol.* 14, 73–84. doi: 10.1080/00223980.1942.9917111

Hou, L., de la Torre, J., and Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: application of the Wald test to investigate DIF in the DINA Model. *J. Educ. Meas.* 51, 98–125. doi: 10.1111/jedm.12036

Huang, G. Y., Zhang, B. S., Wu, Y. Y., Ye, Y. X., and Zhan, J. Z. (2004). *Adolescent Depression Emotion Self-Assessment Scale*. Taipei: Dong's Foundation.

Jaeger, J., Tatsuoka, C., Berns, S. M., and Varadi, F. (2006). Distinguishing neurocognitive functions using partially ordered classification models. *Schizophrenia Bull.* 32, 679–691. doi: 10.1093/schbul/sbj038

Junker, B. M., and Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Appl. Psychol. Meas.* 25, 258–272. doi: 10.1177/01466210122032064

Koenig, H. G., Cohen, H. J., Blazer, D. G., Meador, K. G., and Westlund, R. (1992). A brief depression scale for use in the medically ill. *Int. J. Psychiatry Med.* 22:183.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* 16:606. doi: 10.1046/j.1525-1497.2001.016009606.x

Ma, W., and de la Torre, J. (2016). *GDINA: The Generalized DINA model framework. R Package GDINA (version 0.9.9.8)*.

Ma, W., Iaconangelo, C., and de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Appl. Psychol. Meas.* 40, 200–217. doi: 10.1177/0146621615621717

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika* 64, 187–212. doi: 10.1007/bf02294535

Orlando, M., and Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Appl. Psychol. Meas.* 24, 50–64. doi: 10.1177/01466216000241003

Philipp, M., Strobl, C., de la Torre, J., and Zeileis, A. (2017). On the estimation of standard errors in cognitive diagnosis models. *J. Educ. Behav. Stat.* 43, 88–115. doi: 10.3102/1076998617719728

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Meas.* 1, 385–401. doi: 10.1177/014662167700100306

Rupp, A. A., Templin, J. L., and Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: The Gilford Press.

Smiits, N., Cuijpers, P., and van Straten, Q. (2011). Applying computerized adaptive testing to the CES-D scale: a simulation study. *Psychiatry Res.* 188, 147–155. doi: 10.1016/j.psychres.2010.12.001

Templin, J., and Bradshaw, L. (2014). Hierarchical diagnostic classification models: a family of models for estimating and testing attribute hierarchies. *Psychometrika* 79, 317–339. doi: 10.1007/s11336-013-9362-0

Templin, J. L., and Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *J. Class.* 30, 251–275. doi: 10.1167/iovs.10-5468

Templin, J. L., and Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychol. Methods* 11, 287–305. doi: 10.1037/1082-989x.11.3.287

Tu, D. B., Gao, X. L., Wang, D. X., and Cai, Y. (2017). A new measurement of internet addiction using diagnostic classification models. *Front. Psychol.* 8:1–9. doi: 10.3389/fpsyg.2017.01768

Wang, C., Chang, H. H., and Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *J. Educ. Meas.* 48, 255–273. doi: 10.1111/j.1745-3984.2011.00145.x

World Health Organization [WHO] (2010). *The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines*. Geneva: World Health Organization.

Zung, W. W. (1965). A Self-rating depression scale. *Arch. Gen. Psychiatry* 12, 63–70.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 June 2019
doi: 10.3389/fpsyg.2019.01156

Psychometric Properties and Criterion Validity of STEU-B and STEM-B in Chinese Context

Shuqun Yan[1,2]*, Yuting Feng*[1,2]*, Yaoshan Xu*[1,2]** and Yongjuan Li*[1,2]

[1] *CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, China,*
[2] *Department of Psychology, University of Chinese Academy of Sciences, Beijing, China*

Emotional intelligence (EI) has attracted increasing attention in organizational psychology. The aim of this study was to test the applicability of two performance-based emotional intelligence tests developed in western countries, namely, the brief versions of the Situational Test of Emotional Understanding (STEU-B) and the Situational Test of Emotional Management (STEM-B), in a sample of 904 Chinese employees. Specifically, item response theory (IRT) analyses were conducted. The item parameters along with the item and test information functions of the Chinese versions of the STEU-B and STEM-B were estimated. Moreover, the associations between the STEU-B and STEM-B scores and several work-related variables were examined. The results showed that the STEU-B and STEM-B had acceptable internal consistencies, and similar mean proportions of correct responses, item parameters, item information functions, and test information functions in China, as reported in previous studies. Furthermore, the scores were found to be related to the employees' psychological strain, job-related affect, job satisfaction, and supervisor-rated job performance in a theoretically hypothesized manner. These findings suggested that the STEU-B and STEM-B might be useful measurements in future EI studies in the Chinese organizational context.

Keywords: emotional understanding, emotional management, situational judgment test, item response theory, criterion validity

OPEN ACCESS

Edited by:
Elisa Pedroli,
Istituto Auxologico Italiano (IRCCS),
Italy

Reviewed by:
Ronald H. Humphrey,
Lancaster University, United Kingdom
Ana Altaras Dimitrijevic,
University of Belgrade, Serbia

****Correspondence:***
Yaoshan Xu
xuys@psych.ac.cn

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *03 January 2019*
Accepted: *02 May 2019*
Published: *06 June 2019*

Citation:
Yan S, Feng Y, Xu Y and Li Y
(2019) Psychometric Properties
and Criterion Validity of STEU-B
and STEM-B in Chinese Context.
Front. Psychol. 10:1156.
doi: 10.3389/fpsyg.2019.01156

INTRODUCTION

There is a growing interest in emotional intelligence (EI) in social and organizational psychology, and an increasing number of empirical studies have focused on the criterion validity of EI in predicting real-life outcomes. The EI label has been historically applied to two relatively distinct theoretical constructs: ability EI and trait EI. Ability EI refers to "the ability to perceive emotions, to access and generate emotions so as to assist thought, to understand emotions and emotional knowledge, and to reflectively regulate emotions so as to promote emotional and intellectual growth," which emphasizes EI as an actual ability (Mayer and Salovey, 1997). Trait EI refers to self-perceived emotionality and emotional efficacy that is located within the personality domain (Kafetsios and Zampetakis, 2008). There is evidence of the criterion validity of both ability and trait EI. Ability and trait EI have been found to play important roles in stress management and adaptive coping (Ciarrochi et al., 2002; Oginska-Bulik, 2005), interpersonal relationships and social networks (Brackett et al., 2006; Gallagher and Vella-Brodrick, 2008), intimate relationships (Brackett et al., 2005), and academic achievement (Van Rooy and Viswesvaran, 2004). In the workplace, employees with a high degree of trait EI have been shown to experience more positive and less negative affect (Kafetsios and Zampetakis, 2008), to be more satisfied with their jobs (Kafetsios and Zampetakis, 2008; Greenidge et al., 2014; Meisler, 2014), and to exhibit better job performance (Greenidge et al., 2014; Mulki et al., 2015). A meta-analysis also found that ability EI and trait EI were positively correlated with job performance (O'Boyle et al., 2011). Moreover, empirical evidence revealed that both ability EI and trait EI could act as buffers between job stressors and psychological health (Ciarrochi et al., 2002).

In line with the above definitions, the measurements methods of the two forms of EI are different. Ability EI is assessed through performance-based measurements resembling standard intelligence tests, in which respondents are instructed to maximize effort to achieve the maximum performance on problems related to emotional abilities (Côté, 2014). Trait EI is measured by self-report instruments, through which respondents are asked to confidentially evaluate the contents that describe their abilities in the emotional domain (Schutte et al., 1998). The accuracy of the responses to the self-reported EI items depends on whether the respondents are able to accurately estimate their abilities related to emotional processes and whether they are willing to report them (Côté, 2014). However, evidence has shown that individuals may overestimate their EI (Brackett et al., 2006; Sheldon et al., 2014). Moreover, the self-reported EI questionnaires are susceptible to social desirability bias. For example, applicants may fake their trait EI in these questionnaires during personnel selection. Therefore, EI researchers encourage the use of performance-based measurements to capture actual EI abilities in research and practice, especially in organizational settings (Côté, 2014). Thus, the current study mainly focused on ability EI.

The most prevalent theoretical model in the ability EI research domain is the hierarchical four-branch model, which proposes four branches of ability EI: perceiving/expressing emotions (i.e., accurate perception and expression of emotions); using emotions (i.e., capitalizing on the systematic effects of emotions on cognitive activities); understanding emotions (i.e., identifying the connections between emotions and events); and regulating emotions (i.e., increasing, maintaining, or decreasing one's own or others' emotions) (Mayer and Salovey, 1997). Based on this model, Mayer et al. (2002) developed the Mayer-Salovey-Caruso Emotional Intelligence Test (MSCEIT) to measure these four EI branches. To date, research on ability EI has been dominated by the MSCEIT, and thus, what we know about ability EI is largely based on this measurement. However, it is difficult to know whether these empirical results were attributable to the constructs examined or the unique measurement method used. Moreover, there is evidence to suggest that the MSCEIT has problems with its scoring method (Austin et al., 2008), as well as with its task and item selection (Roberts et al., 2006), which emphasizes the necessity and importance of developing alternative measures of ability EI.

To provide alternative instruments for assessing ability EI, MacCann and Roberts (2008) developed the situational test of emotional understanding (STEU) and the situational test of emotional management (STEM) using the situational judgment test paradigm. The STEU and STEM target the third and the fourth branch of the four-branch ability EI model, respectively. According to this model, the four hierarchically ordered EI branches monotonically increase in cognitive complexity from the first to fourth branch, and can be grouped into two areas: experiential EI (encompassing the lower two branches) and strategic EI (encompassing the two higher branches) (Mayer et al., 2002). Thus, the STEU and STEM provide a comprehensive picture of strategic EI. The understanding emotions branch is the "most cognitively saturated" and regarded as the key focus of abstract processing and reasoning with respect to emotion (Mayer et al., 2001). The regulating emotions branch is the highest and most complex branch; it involves managing emotions for personal and interpersonal growth, which combines and balances motivational, emotional, and cognitive factors (Mayer et al., 2001). A recent empirical study indicated that the discriminating and predictive power of ability EI lay primarily in these two strategic branches (Dimitrijević et al., 2018).

The STEU measures an individuals' ability to understand the connections between events and emotions (i.e., the understanding emotions branch) (MacCann and Roberts, 2008). The content of the items of the STEU was derived from Roseman (2001) appraisal theory, which provided a strong theoretical basis for emotional understanding. Within the framework of this theory, individuals' evaluation of a situation or event cause specific reactions and bring about emotional responses based on their appraisal, and 17 discrete emotions are generated according to specific combinations of seven appraisal dimensions (motive-consistency, causal attribution, certainty, control potential, unexpectedness, motivational state, and problem source). The STEU consists of 42 scenarios covering the following emotions: sadness, pride, relief, joy, regret, gratitude, distress, hope, contempt, surprise, frustration, anger, fear, and dislike. The scenarios contain ample multiple-choice items, including 14 context-reduced items, 14 with a personal-life context, and 14 with a workplace context (MacCann and Roberts, 2008). In each scenario, an emotional situation is described, and five emotions are presented. Respondents are asked to indicate which emotion is most likely to be generated by that particular situation. The answers of the items are scored as either correct or incorrect based on the appraisal theory. Thus, the scoring system of STEU is theoretically based and substantially different from the scoring system used for the MSCEIT. The STEM measures individuals' ability to cope with stressful events by regulating negative emotions and enhancing positive emotions through emotional management (i.e., the regulating emotions branch), which is developed on the basis of the situational judgment test paradigm. In accordance with this paradigm, items were generated by the semi-structured interviews, and answers from participants about those items constituted the response options. The relevant experts decided the scoring system based on their selection for the proportion of each option (MacCann and Roberts, 2008). The test consists of 44 scenarios covering three emotions, namely, fear, anger, and sadness. In each scenario, an emotional situation is described and four options regarding the action to manage the emotions and solve the problems in that scenario are presented. The respondents are asked to select the most effective option. The STEU and STEM showed good convergent and divergent validity. The correlation between the STEU and STEM scores was 0.29 (Austin, 2010). The STEU scores correlated at 0.44 with the MSCEIT understanding scores (Austin, 2010) and at 0.31 with scores on the theory of mind test (Ferguson and Austin, 2010). The STEM scores correlated at 0.30 with the MSCEIT management scores (Austin, 2010) and at 0.21 with scores on the theory of mind test (Ferguson and Austin, 2010). The STEU and STEM also showed small to moderate correlations with personality traits (MacCann and Roberts, 2008; Libbrecht and Lievens, 2012). Moreover, the STEM scores correlated at 0.23 with academic performance in a sample of undergraduate medical students (Libbrecht et al., 2014), thus providing support for criterion validity in real life.

More recently, researchers have developed the brief version of STEU (STEU-B) and the brief version of STEM (STEM-B) by evaluating the psychometric properties of STEU and STEM using the item response theory (IRT) method (Allen et al., 2014, 2015). IRT provides valuable methods for assessing the psychometric properties of EI measurements (Karim, 2010; Cho et al., 2015), which has advantages compared with the classical test theory (CTT) method. First, unlike CTT, which examines the psychometric properties of EI measurements based on observed scores, the IRT method provides psychometric information that is not dependent on the sample. Furthermore, the CTT method assumes a constant effectiveness and measurement precision of the test and items. In comparison, the IRT method holds that the effectiveness and precision of the test and items vary across different levels of the trait. Therefore, the IRT can be used to calculate the probability that the respondents choose a particular answer of each item and to estimate the ability of the test and each item to differentiate respondents at every level of EI. Allen and colleagues evaluated the item parameters (i.e., discrimination, difficulty, and guessing parameters) and the item information for each item included in STEU and STEM as provided by IRT analysis (Allen et al., 2014, 2015). Based on these psychometric properties, the items with low "maximum effectiveness" (a maximum amount of item information < 0.05) and providing information for similar areas of the latent scale were omitted, resulting in 19-item STEU and 18-item STEM scales. Thus, the STEU-B and STEM-B can provide sufficient information across different levels of item difficulty. The Cronbach's alpha coefficients for STEU-B and STEM-B were 0.63 (Allen et al., 2014) and 0.84 (Allen et al., 2015), respectively. The correlation between STEU-B and STEM-B was 0.30 (Allen et al., 2015). With the increasingly high usage of EI measurements in research and practice, the short version of performance-based EI instruments has been requested by both EI researchers and organizational managements. Thus, STEU-B and STEM-B can be useful tools in cases where research time is limited and for organizational management purposes.

Despite these significant advances in EI research, STEU and STEM research has been limited to Western cultural participants (e.g., MacCann and Roberts, 2008; Austin, 2010; Côté et al., 2011). Previous evidence has indicated that cultural differences between performance-based EI tests may exist (Côté, 2014). Therefore, the generalization of STEU-B and STEM-B should be further examined in different cultural contexts. Furthermore, EI is an increasingly important issue in the workplace setting, and applying EI instruments in organizational management comes with the growing need to evaluate the measurement precision and criterion validity of EI instruments in the organizational setting (Karim, 2010; Greenidge et al., 2014). However, empirical evidence for the criterion validity of STEU-B and STEM-B to predict the work-related variables in a real organizational setting is limited. It is also unknown whether the patterns of associations between EI and work criteria that have been found in research on Western culture hold in Chinese organizational context. Accordingly, this study aimed to validate the STEU-B and STEM-B in a sample of Chinese employees in terms of psychometric properties and criterion validity. Specifically, we analyzed the psychometric properties of the Chinese versions of STEU-B and STEM-B using the IRT method and examined the associations between the Chinese versions of STEU-B and STEM-B scores and several work-related variables. By doing so, this study improved the research on EI in different cultural contexts and extended the information on the STEU-B and STEM-B by providing their criterion-related validity in the Chinese organizational setting.

Specifically, we expected the Chinese versions of the STEU-B and STEM-B scores to be related to the work-related criterion in several respects. First, we posited that EI scores should be negatively associated with the indicators of occupational stress and strain. The abilities of emotional understanding and emotional regulation facilitate stress management and adaptive coping (Oginska-Bulik, 2005; Gallagher and Vella-Brodrick, 2008). Thus, employees who are capable of understanding and regulating emotions can cope with negative events and occupational stress well, and thereby suffer less psychological strain than employees with low EI levels. Second, EI should be related to positive and negative affect at work. To be specific, the emotional regulation branch of EI can help employees to cope with high job demands and undesirable job-related events, as well as to control and alter emotional experiences caused by unfavorable events, which may lead to more positive experiences and less negative experiences at work. Consistent with this reasoning, evidence has shown that employees with high ability of emotional regulation experienced more work-related positive affect and less work-related negative affect than employees with low emotional regulation ability (Kafetsios and Zampetakis, 2008; Parke et al., 2015). Third, we posited that emotionally intelligent employees should be more satisfied with their jobs. Employees with high abilities of emotional understanding and regulation can better understand and anticipate others' emotions in the workplace, cope with negative experiences and unfavorable job-related events, and have better psychological health than others (Sy et al., 2006; Vratskikh et al., 2016). This can in turn increase their job satisfaction levels. Existing research has consistently suggested that EI predicts employees' job satisfaction (Kafetsios and Zampetakis, 2008; Greenidge et al., 2014; Ouyang et al., 2015; Vratskikh et al., 2016). Thus, the STEU-B and STEM-B scores should be positively associated with employees' job satisfaction. Fourth, EI is an important predictor of job performance. In two meta-analyses, the correlations between performance-based EI scores and job performance were 0.16 (Joseph and Newman,

2010) and 0.21 (O'Boyle et al., 2011), respectively. Moreover, emotional understanding and emotional regulation were found to play different roles in determining job performance. In the cascading model of EI (Joseph and Newman, 2010; Newman et al., 2010), emotional understanding was proposed as an effect on emotional regulation, which in turn influenced job performance directly. Therefore, emotional regulation mediated the effect of emotional understanding on job performance. Accordingly, STEU-B and STEM-B scores should be positively related to job performance. Moreover, the association between STEM-B score and job performance should be stronger, and the effect of STEU-B score on job performance would be fully mediated by the STEM-B score.

In summary, based on the existing results of STEU and STEM, and the research on EI in the organizational context, the following hypotheses were proposed: (1) A significant correlation exists between the Chinese versions of STEU-B and STEM-B scores; (2) The Chinese versions of the STEU-B and STEM-B scores are negatively correlated with psychological strain; (3) The Chinese versions of the STEM-B score are positively correlated with positive affect at work and negatively correlated with negative affect at work; (4) The Chinese versions of the STEU-B and STEM-B scores are positively correlated with job satisfaction; and (5) The Chinese versions of the STEU-B and STEM-B scores are positively correlated with job performance, and the effect of the STEU-B score on job performance is fully mediated by the STEM-B score.

MATERIALS AND METHODS

Participants and Procedures

The sample for this research was drawn from full-time employees working in an information technology company located in three major cities (Beijing, Shanghai, and Guangzhou) of China. Before the study, we contacted the company's human resource management department to help us to distribute the survey. The employees were invited to participate in the study voluntarily. Participants went to a meeting room during their break time and were briefed on the purpose and procedure of the current study by a researcher individually. They were also assured that their responses would be kept anonymous and confidential. Then each participant provided written, informed consent prior to data collection. After that, they were asked to complete the STEU-B, STEM-B, and to participate in the measurement of criterion-related variables individually in the meeting room. In total, 904 participants completed and returned the survey. The sample consisted of 537 men and 367 women with an average age of 27.72 years (*SD* = 3.30) and an average job tenure in the current company of 4.20 years (*SD* = 2.52). The education level of the sample was relatively high; 25 participants (3.2%) had a high school diploma, 648 participants (82.4%) had a college education, and 113 participants (14.4%) had a master's degree. The employees were from various departments: marketing and sales (20.0%), technology and data analysis (22.5%), product development (12.4%), customer service and consulting (13.8%), administration (12.5%), human resources (3.5%), finances (6.7%), and unspecified other departments (8.6%). Among these employees, 378 (41.8%) needed to interact with customers (e.g., sales, customer service technicians, and product managers), 438 (48.5%) required frequent team discussion and cooperation (e.g., consultors, product managers, and products technicians), and 85 (9.4%) were team leaders. The direct supervisors of the participants were invited to confidentially evaluate the job performance of their subordinates. We received 632 supervisor evaluations.

All of the procedures performed in studies involving human participants were approved by The Ethics Committee of the Institute of Psychology of the Chinese Academy of Sciences. Approval of the study was also done by the human resource management department of the company at which this study was conducted.

Measures

The STEU-B and the STEM-B are described in the Introduction section. The English versions of the STEU-B and STEM-B were translated and adapted to the Chinese language in several stages. First, the original English versions were translated into Chinese by three bilingual native Chinese researchers independently. This resulted in different initial versions, which were reviewed and compared to produce consensual versions of STEU-B and STEM-B by the authors of the present study. Then, another bilingual native Chinese researcher back-translated these Chinese versions into English. The backward translator was familiar with the Chinese and western cultures and had no access to the original English versions. Next, the back-translated English versions were compared with the original English versions. Items with problematic back translations were thoroughly discussed by the authors and other experts in the field of emotion through a series of group meetings, and some minor revision were made to ensure the culture equivalence between the original English versions and the Chinese versions. Most modifications were minor, involving the choice between two synonyms or the change of the word order. The STEU-B was scored according to the original scoring system. Specifically, the correct answer was scored as "1" and the other answers were scored as "0" (MacCann and Roberts, 2008). The STEM-B scoring system is ordinarily based on the experts' proportion of choosing each answer (MacCann and Roberts, 2008). In this study, we used the dichotomous scoring suggested by Allen et al. (2015) so that the IRT analyses could be conducted. Specifically, the best option was scored as "1," and the other answers were scored as "0."

The Chinese version of the General Health Questionnaire (GHQ-12) (Wang and Lin, 2011) was used to measure the psychological strain of employees. The questionnaire consisted of 12 items. Participants evaluated the levels of their psychological strain on a 7-point Likert scale (from 1 = strongly disagree to 7 = strongly agree), with higher scores indicating a higher level of psychological strain.

The IWP Multi-Affect Indicator (Warr et al., 2014) revised by Li et al. (2017) in the Chinese organizational context was used to assess participants' experience of work-related positive and negative affect. This scale defined affect at work into four states: high-activation pleasant affect (HAPA), low-activation pleasant affect (LAPA), high-activation unpleasant affect (HAUA), and low-activation unpleasant affect (LAUA). Each dimension was measured using four adjectives that described work-related affect (HAPA: being enthusiastic, excited, inspired, and joyful; LAPA: being at ease, calm, laid back, and relaxed; HAUA: being anxious, nervous, tense, and worried; and LAUA: being dejected, depressed, despondent, and hopeless). The participants rated their experience at work in the past 4 weeks on a 7-point Likert scale (from 0 = never to 6 = always). As recommended by Warr and Parker (2010), the four single-quadrant scores were combined to create four double-quadrant dimensions: the positive affect dimension (all pleasant affect items) with higher scores indicating a higher level of positive affect, the negative affect dimension (all unpleasant affect items) with higher scores indicating a higher level of negative affect, the anxiety-comfort dimension (LAPA and reverse-scored of HAUA) with higher scores indicating a higher level of comfort, and the depression-enthusiasm dimension (HAPA and reverse-scored of LAUA) with higher scores indicating higher level of enthusiasm.

The job satisfaction scale developed by Schriesheim and Tsui (1980) was also employed. The scale consisted of 6 items. Respondents indicated their satisfaction with different aspects of their current job (e.g., co-workers, supervisors, and promotion) on a 5-point Likert scale (from 1 = very unsatisfied to 5 = very satisfied).

Supervisors were then asked to evaluate the general job performance of their subordinate on a 4-point scale (1 = fails, 2 = needs improvement, 3 = succeeds/meets standards, 4 = excels/exceeds standards). This measurement originated from Leavitt et al. (2011).

Demographic data on the employees (i.e., gender, age, and job tenure) were collected as control variables.

Data Analysis Procedure

Descriptive statistics (mean scores and standard deviation), item-total score correlation indexes, and Cronbach's alpha coefficients were computed.

Before the IRT analysis, the unidimensionality of the scale had to be examined because IRT assumes that the items included in the scale assess a single construct. Therefore, confirmatory factor analyses (CFA) were conducted to verify the unidimensionality of the STEU-B and STEM-B data.

IRT analyses were then conducted using the latent trait modeling package of R software (Rizopoulos, 2006). According to the dichotomous nature of the data, the 3-parameter logistic (3-PL) IRT model (Birnbaum, 1968) was used to fit the STEU-B and STEM-B items. With the 3-PL IRT model, the discrimination, difficulty, and guessing parameters were calculated. The discrimination parameters (ai) captured the relationship between the probability of endorsing the correct option for each item and the latent construct, which represented the discriminating power of the particular item. The discrimination parameters were interpreted qualitatively with the Baker (1985) classification using the following terms: $a < 0.20$, very low discrimination; $0.21 < a < 0.40$, low discrimination; $0.41 < a < 0.80$, moderate discrimination; $0.81 < a < 1$, high discrimination; $a > 1$, very high discrimination. The difficulty parameters (bi) indicated the θ value (i.e., the latent trait) at which people had a 50% chance of selecting the correct answer and at which point the item could provide sufficient information. The guessing parameters (ci) represented the index of correct guessing, which reflected the probability of choosing the correct answer.

The item information curve (IIC) for each item was generated based on the IRT parameters, which described the distribution of information provided by an item across the continuum of the latent trait (θ). The area under IIC equaled the amount of information that the particular item could provide across the different levels of the latent trait. The amount of information indicated the ability of the item to distinguish the respondents with different levels of EI. The test information function (TIF) of the scale was calculated by aggregating the IICs of all items within the scale. The area under TIF represented the total test information.

To investigate the criterion-related validity of the Chinese versions of STEU-B and STEM-B, the partial correlations between the STEU-B score, STEM-B score, as along with the psychological strain, job-related effects, job satisfaction, and general job performance by controlling gender, age, and job tenure were calculated. Moreover, since the different effects of the STEU-B and STEM-B scores on job performance were expected, we conducted a hierarchical regression analysis that predicted job performance.

RESULTS

Basic Descriptive Statistics

Tables 1, **2** list the mean score, standard deviation, and correlation between items and the total score for each item within the STEU-B and STEM-B scales, respectively. The mean scores on the STEU-B and STEM-B scales were 0.63 (*SD* = 0.19) and 0.60 (*SD* = 0.21), respectively. The Cronbach's alpha coefficients for the STEU-B and STEM-B were 0.72 and 0.75, respectively. For the 19 STEU-B items, the correlations between items and the total score ranged from 0.33 to 0.54. For the 18 STEM-B items, the correlations between items and the total score ranged from 0.34 to 0.49. A significant gender difference was observed in the scores on the STEM-B (males: $M = 0.58$, $SD = 0.22$, $n = 537$; females: $M = 0.63$, $SD = 0.18$, $n = 367$; $t = -3.15$; $p = 0.002$; Cohen's $d = 0.26$). However, no significant gender difference was observed in the scores on the STEU-B (males: $M = 0.62$, $SD = 0.19$, $n = 537$; females: $M = 0.63$, $SD = 0.18$, $n = 367$; $t = -0.26$; $p > 0.05$; Cohen's $d = 0.06$).

Unidimensionality

In an IRT analysis, ensuring unidimensionality of the measurement is important. Therefore, CFA was conducted to test the unidimensionality of the STEU-B and STEM-B scales. The results showed that the one-factor model fitted the data on the Chinese version of the STEU-B well [$\chi^2 = 232.80$, $df = 152$, $GFI = 0.97$, $CFI = 0.93$, $IFI = 0.93$, RMSEA = 0.024, 90% CI = (0.018, 0.030)]. The fit indices for the STEM-B scale were similar [$\chi^2 = 286.43$, $df = 135$, $GFI = 0.97$, $CFI = 0.90$, *IFI* = 0.90, RMSEA = 0.035, 90% *CI* = (0.030, 0.041)]. These results provided supports for the unidimensionality of the STEU-B and STEM-B.

Item Parameter Estimation and Information

The 3-PL model was used to fit the 19 STEU-B items. **Table 1** shows the item parameters and the information for each item. The discrimination parameters ranged from 0.57 to 1.81, the difficulty parameters ranged from -1.67 to 0.97, and the guessing parameters ranged from 0.01 to 0.13. The item information for each item ranged from 0.43 to 1.44, and the maximum amount of item information ranged from 0.09 to 0.53. The total test information for the STEU-B scale was 14.91, and the point of maximum test information on the θ scale was −0.61, which suggested that the STEU-B scale can provide more sufficient information for individuals with low emotional understanding ability than those with high emotional understanding ability.

The 3-PL model was used to fit the 18 STEM-B items. **Table 2** shows the item parameters and the information for each item. The discrimination parameters ranged from 0.68 to 1.62, the difficulty parameters ranged from -2.00 to 1.00, and the guessing parameters ranged from 0.01 to 0.13. The item information for each item ranged from 0.61 to 1.27, and the maximum amount of item information ranged from 0.13 to 0.45. The test information for the STEM-B scale was 16.27, and the point of maximum test information on the θ scale was −0.42, which suggested that the STEM-B scale can provide more sufficient information for individuals with low emotional management ability than those with high emotional management ability.

Correlations of STEU-B, STEM-B and Criterion Variables

The partial correlations among the STEU-B score, the STEM-B score, and other criterion-related variables by controlling age, gender, and job tenure are shown in **Table 3**. The STEU-B score was significantly correlated with the STEM-B score ($r = 0.32$, $p < 0.001$). The STEU-B score was significantly and negatively correlated with psychological strain, LAUA and overall negative affect at work. It significantly and positively correlated with LAPA, overall positive affect, the anxiety-comfort score, and the depression-enthusiasm score, job satisfaction, and supervisor-rated general job performance. The STEM-B score was significantly associated with all measured criterion-related variables in the expected directions.

Regression Analysis Predicting Job Performance

To further explore the differential predictive power of STEU-B and STEM-B on job performance, a hierarchical regression analysis predicting job performance was conducted. Independent variables and outcome variable were standardized to control the size of the effects. First, gender, age, and job tenure were entered as control variables. Second, the STEU-B score was entered into the regression. The results showed that this score significantly

TABLE 1 | Descriptive statistics, item parameters and item information for STEU-B (*n* = 904).

Item	Descriptive statistics			Item parameters estimates			Information Total Test = 14.91	
	Mean	SD	r_{it}	a_i	b_i	c_i	Information	Information_{max}
1	0.54	0.50	0.42***	0.83	−0.18	0.01	0.78	0.19
2	0.30	0.46	0.40***	0.91	0.97	0.05	0.68	0.15
3	0.47	0.50	0.42***	0.91	0.28	0.03	0.78	0.18
4	0.66	0.47	0.54***	1.81	−0.38	0.12	1.44	0.53
5	0.47	0.50	0.39***	0.72	0.24	0.01	0.65	0.14
6	0.73	0.44	0.42***	0.93	−1.22	0.01	0.87	0.22
7	0.75	0.44	0.46***	1.32	−0.87	0.13	1.13	0.34
8	0.84	0.37	0.43***	1.23	−1.67	0.01	1.20	0.39
9	0.73	0.44	0.43***	0.98	−1.18	0.01	0.92	0.24
10	0.47	0.50	0.37***	0.64	0.25	0.01	0.52	0.10
11	0.73	0.44	0.40***	0.84	−1.38	0.01	0.78	0.19
12	0.47	0.50	0.37***	0.83	0.63	0.13	0.53	0.11
13	0.73	0.47	0.40***	0.82	−1.30	0.01	0.78	0.18
14	0.69	0.46	0.33***	0.57	−1.48	0.01	0.43	0.09
15	0.66	0.47	0.42***	0.83	−0.87	0.01	0.77	0.18
16	0.74	0.44	0.36***	0.68	−1.64	0.01	0.60	0.13
17	0.77	0.42	0.39***	0.80	−1.61	0.01	0.76	0.18
18	0.60	0.59	0.43***	0.82	−0.54	0.01	0.75	0.17
19	0.57	0.49	0.37***	0.67	−0.39	0.03	0.54	0.12

*STEU-B, brief version of Situational Test of Emotional Understanding; SD, standard deviation; r_{it}, item-total correlation; a_i, discrimination parameter; b_i, difficulty parameter; c_i, guessing parameter; Information_{max}, maximum amount of item information. ***$p < 0.001$.*

TABLE 2 | Descriptive statistics, item parameters and item information for STEM-B (*n* = 904).

Item	Descriptive statistics			Item parameters estimates			Information Total Test = 16.27	
	Mean	**SD**	**r_{it}**	**a_i**	**b_i**	**c_i**	**Information**	**$\text{Information}_{\max}$**
1	0.47	0.50	0.43***	0.89	0.22	0.01	0.79	0.19
2	0.54	0.50	0.43***	0.83	−0.14	0.01	0.75	0.17
3	0.45	0.50	0.47***	1.53	0.54	0.13	0.94	0.25
4	0.59	0.49	0.47***	1.01	−0.39	0.01	0.97	0.26
5	0.67	0.47	0.43***	0.87	−0.90	0.01	0.86	0.21
6	0.71	0.45	0.41***	0.84	−1.16	0.01	0.82	0.20
7	0.62	0.49	0.42***	0.82	−0.62	0.01	0.76	0.17
8	0.69	0.46	0.44***	0.94	−0.91	0.01	0.94	0.24
9	0.73	0.44	0.46***	1.09	−1.08	0.01	1.14	0.35
10	0.55	0.50	0.47***	1.03	−0.17	0.01	0.96	0.26
11	0.64	0.48	0.49***	1.16	−0.55	0.01	1.12	0.34
12	0.58	0.49	0.42***	0.95	−0.07	0.12	0.77	0.18
13	0.48	0.50	0.49***	1.62	0.39	0.12	1.01	0.28
14	0.63	0.48	0.46***	0.96	−0.57	0.01	0.94	0.24
15	0.78	0.42	0.34***	0.68	−2.00	0.01	0.61	0.13
16	0.36	0.48	0.41***	1.31	1.00	0.12	0.71	0.16
17	0.50	0.50	0.40***	0.79	0.08	0.01	0.69	0.15
18	0.83	0.38	0.46***	1.23	−1.57	0.01	1.27	0.45

*STEM-B, brief version of Situational Test of Emotional Management; SD, standard deviation; r_{it}, item-total correlation; a_i, discrimination parameter; b_i, difficulty parameter; c_i, guessing parameter; Information_{max}, maximum amount of item information. ***$p < 0.001$.*

predicted job performance (β = 0.11, *p* = 0.008). Third, the STEM-B score was entered. The results showed that the STEM-B score significantly predicted job performance (β = 0.20, *p* < 0.001), whereas the coefficient of the STEU-B score became insignificant (β = 0.04, *p* > 0.05). Moreover, bootstrap results suggested that the standardized coefficient for the indirect effect of the STEU-B score on job performance through the STEM-B score was significant [effect = 0.07; 95% *CI* = (0.37, 0.11)].

DISCUSSION

This study examined the psychometric properties of STEU-B and STEM-B using the IRT method and their criterion validity in a sample of 904 Chinese employees. The internal consistencies of the Chinese versions of the STEU-B and STEM-B scales were found to be adequate; both were above 0.70. The mean scores on STEU-B and STEM-B in the Chinese context were close to those on the original version in the Western context (Allen et al., 2014, 2015). Previous studies reported that east Asians performed worse on MSCEIT than did North Americans (Mayer et al., 2002). This cultural difference in the scores on the performance-based EI test was in part because the test was developed in the west, and the correct answers to problems about emotions in the test varied across different cultures (Moon, 2011; Côté, 2014). However, our results indicated that the correct answers and scoring systems of STEU-B and STEM-B that were developed in the west were also applicable in the Chinese context.

Furthermore, the IRT analyses revealed that all of the items within the original STEU-B and STEM-B scales had good discrimination parameters in the Chinese context (moderate to high level). Moreover, the difficulty values of these items were evenly spaced, ranging from −2.00 to 1.00. The item information for each item was then computed as a function of item parameters. The maximum amount of item information ranged from 0.09 to 0.53 in this study, which exceeded the cutoff value of 0.05 suggested by Allen et al. (2014). These results were in line with previous findings, which showed that the items included in the STEU-B and STEM-B were able to distinguish different levels of EI effectively and provide sufficient item information (Allen et al., 2014, 2015). The inspection of both the IIFs and TIFs revealed that the Chinese versions of STEU-B and STEM-B had uneven information functions, and that STEU-B and STEM-B provided the maximum information for individuals with a trait value of −0.61 and a trait value of −0.42, respectively. Thus, similar to the English version, the Chinese versions of STEU-B and STEM-B were proved to be more useful in identifying individuals with poor to average emotional understanding and emotional management (Allen et al., 2014, 2015). Taken together, these results indicated that the psychometric properties of the Chinese versions of STEU-B and STEM-B were satisfactory, and that the original scoring systems of these scales were applicable in the Chinese context.

The criterion validity of the Chinese versions of STEU-B and STEM-B was evaluated by determining whether the STEU-B and STEM-B scores were related to several work criteria in meaningful ways. Consistent with substantial EI research reported in the west, which suggested that EI played an important role in stress management and job satisfaction (Oginska-Bulik, 2005; Gallagher and Vella-Brodrick, 2008; Vratskikh et al., 2016),

TABLE 3 | Descriptive statistics and partial correlations of all variables.

	Mean	SD	1	2	3	4	5	6	7	8	9	10	11	12
1. STEU-B	0.63	0.19	*0.72*											
2. STEM-B	0.60	0.21	0.32***	*0.75*										
3. Psychological strain	2.57	0.66	−0.19***	−0.22***	*0.87*									
4. HAPA	4.17	1.02	0.06	0.25***	−0.46***	*0.87*								
5. LAPA	4.17	1.04	0.12***	0.23***	−0.46***	0.60***	*0.85*							
6. HAUA	2.90	0.81	−0.06	−0.20***	0.50***	−0.30***	−0.36***	*0.82*						
7. LAUA	2.34	0.73	−0.19***	−0.24***	0.59***	−0.32***	−0.29***	0.61***	*0.83*					
8. Positive affect	4.17	0.92	0.10**	0.27***	−0.51***	0.89***	0.89***	−0.37***	−0.34***	*0.90*				
9. Negative affect	2.62	0.69	−0.14***	−0.25***	0.60***	−0.35***	−0.36***	0.91***	0.88***	−0.40***	*0.88*			
10. AC	4.13	0.76	0.11**	0.27***	−0.57***	0.57***	0.87***	−0.77***	−0.52***	0.81***	−0.73	*0.84*		
11. DE	4,41	0.72	0.14***	0.31***	−0.62***	0.88***	0.58***	−0.53***	−0.74***	0.81***	−0.70***	0.67***	*0.84*	
12. Job satisfaction	3.62	0.63	0.22***	0.30***	−0.37***	0.38***	0.25***	−0.22***	−0.34***	−0.31***	0.29***	0.44***	−0.31***	*0.81*
13. Job performance	3.05	0.67	0.11**	0.21***	−0.31***	0.20***	0.24***	−0.22***	−0.20***	−0.24***	0.28***	0.25***	−0.24***	0.17***

*STEU-B, brief version of Situational Test of Emotional Understanding; STEM-B, brief version of Situational Test of Emotional Management; HAPA, high-activation pleasant affect; LAPA, low-activation pleasant affect; HAUA, high-activation unpleasant affect; LAUA, low-activation unpleasant affect; AC, anxiety-comfort; DE, depression-enthusiasm. Diagonal italic numbers represent the internal consistency coefficient of the scale. n = 632 for correlations with job performance, n = 904 for other correlations. **p < 0.01, ***p < 0.001.*

the Chinese versions of the STEU-B and STEM-B scores were significantly related to a reduction in employees' psychological strain and an increase in their job satisfaction. The results also demonstrated that the STEM-B score had positive relationships with both the HAPA and LAPA, and negative relationships with both the HAUA and LAUA at work, whereas the STEU-B score was only weakly associated with LAPA (e.g., being at ease) and LAUA (e.g., feeling dejected). These results were in line with previous studies that suggested that regulation of emotion was a more predictive EI dimension of work-related effects than of emotional understanding (Kafetsios and Zampetakis, 2008; Parke et al., 2015). Although the relationships between the STEU-B score and work-related effects were not expected, these results indicated that the employees with a high degree of emotional understanding experienced lower levels of LAUA in the Chinese organizational context. Both STEU-B and STEM-B were also significantly associated with double-quadrant dimensions affective scores. However, the correlations between STEU-B and these scores were very weak. Overall, the observed correlations between STEU-B and criteria indicated that the STEU-B had a stronger correlation with job satisfaction which involved the cognitive evaluation regarding different aspects of work, whereas the associations between STEU-B and affect-related scores were weaker. These results were consistent with the theoretical argument that emotional understanding was the most "cognitive" EI branch, which had a strong association with abstract reasoning and emotional information-processing (Mayer et al., 2001).

Our results also demonstrated that both the STEU-B and STEM-B scores were related to the supervisor-rated general job performance, and the association between STEM-B score and job performance was stronger than that between STEU-B score and job performance. The correlations in this study were similar to those reported in previous meta-analyses (Joseph and Newman, 2010; O'Boyle et al., 2011). The regulating emotions branch is the highest and most complex EI branch, which involves motivational, emotional, and cognitive factors. Thus, it may facilitate employees' general job performance by achieving more adaptive mood states, obtaining valuable resources, forming better relationships with coworkers or customers, and promoting personal growth. Furthermore, in line with the cascading model of EI, which proposed that the higher branches of abilities (e.g., emotional regulation) were developed on the basis of the lower branches of abilities (e.g., emotional understanding) (Joseph and Newman, 2010; Newman et al., 2010), our results indicated that the understanding of emotions in specific situations may impact the management of emotions, such as the strategies we use to regulate our emotions, which in turn contribute to job performance. The practical implication of this is that it is meaningful to utilize some training programs to improve the emotional understanding of ability EI before emotional regulation to enhance employee's job performance.

The STEU-B and STEM-B target the two higher, strategic branches of the ability EI that are important in the organizational context. The STEU-B and STEM-B are theoretically based and provide sufficient test information with fewer items, which is time-saving. Therefore, it would be very useful when testing time is severely limited and for researches that focus on strategic EI rather than experiential EI. Moreover, unlike MSCEIT which is a commercial test with scoring performed by a test company, the items selection and scoring systems of STEU-B and STEM-B are provided clearly to EI researchers. Thus, it is possible to further develop and improve these instruments. However, there are some limitations to this method of measurement. The item selection was based on test information curves, and this might have decreased the measurement precision for respondents whose ability lay outside of the mean (Allen et al., 2015). The mean scores of the Chinese versions of STEU-B and STEM-B in the current study were also found to be higher than those of the original full-length versions (MacCann and Roberts, 2008), indicating that the easier items were selected. Thus, the STEU-B and STEM-B would be more useful in populations where lower levels of emotional understanding and management are assumed.

Some limitations of this study and directions for further research should be addressed. First, the sample of this study was derived from a high-tech organization in three major cities of China, where the level of educational attainment was relatively high. In addition, these participants were relatively young, and different patterns in EI may be affected by individuals' growth. Therefore, future studies should include broader samples of different occupations, education levels, socioeconomic backgrounds, and age groups to generalize these measurements. Second, although we provided evidence for the criterion validity of STEU-B and STEM-B in a Chinese organizational setting by examining their relationships with several important work-related criteria, the incremental validity was not examined since we did not control other individual different variables that predicted work-related criteria, such as cognitive ability, personality traits, and self-reported EI (Joseph and Newman, 2010; O'Boyle et al., 2011). Recent meta-analysis studies provided support for the incremental validity of EI in predicting work attitude (Miao et al., 2017) and job performance (Miao et al., 2018) while controlling for the big five personality traits and cognitive ability. Therefore, it is of great importance to explore the incremental validity of STEU-B and STEM-B in the organizational context by including these variables. Third, the associations between ability EI and general job performance were proved to be weak in our study. It has been proposed that the relationships between EI and work outcomes depend on the job or employment setting. Thus, considering other variables that is related to the specific work situation, or other work criteria are also important. For example, further studies can relate the STEU-B and STEM-B to other work criteria, such as emotional labor, contextual performance, and leadership. Fourth, the underlying cognitive processes may be different for different format (multiple-choice or rate-the-extent) (MacCann and Roberts, 2008), thus future research could explore the influence of thinking mode in EI research. Finally, we did not consider the influence of cultural values on EI and work-related outcomes, such as collectivism and long-term orientation (Miao et al., 2018). Researchers should incorporate these factors when delving into this topic in the future.

CONCLUSION

This study examined the applicability of two performance-based EI tests, namely STEU-B and STEM-B, in a sample of 904 Chinese employees. The internal consistencies were acceptable. The item parameters provided by the IRT analyses showed good discriminatory power and reasonable variation in difficulty across all the items within the STEU-B and STEM-B scales. Moreover, the scores on STEU-B and STEM-B were associated with several emotion- and work-related criteria in meaningful ways. Taken together, the Chinese versions of STEU-B and STEM-B scales were found to be psychometrically adequate measurements which might be useful to capture employees' emotional understanding and emotional regulation as alternative ability EI tests. Further research should focus on further validation in broader work contexts, and in relation with various personality traits, intelligence, and work-related outcomes.

ETHICS STATEMENT

All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

AUTHOR CONTRIBUTIONS

SY collected the data, analyzed and interpreted the data, wrote the manuscript, and was involved in the study conception and design. YF collected the data, analyzed and interpreted the data, and was involved in manuscript preparation and revision. YX conceived and designed the study, analyzed and interpreted the data, reviewed and edited the manuscript, and provided final approval of the version. YL conceived and designed the study, analyzed and interpreted the data, and was involved in manuscript preparation.

FUNDING

This study was partly supported by the National Natural Science Foundation of China (Grant No. 71501177) and the National Key Research and Development Program of China (Grant No. 2016YFC0802600).

REFERENCES

Allen, V., Rahman, N., Weissman, A., MacCann, C., Lewis, C., and Roberts, R. D. (2015). The situational test of emotional management-brief (STEM-B): development and validation using item response theory and latent class analysis. *Pers. Individ. Dif.* 81, 195–200. doi: 10.1016/j.paid.2015.01.053

Allen, V. D., Weissman, A., Hellwig, S., MacCann, C., and Roberts, R. D. (2014). Development of the situational test of emotional understanding-brief (STEU-B) using item response theory. *Pers. Individ. Dif.* 65, 3–7. doi: 10.1016/j.paid.2014.01.051

Austin, E. J. (2010). Measurement of ability emotional intelligence: results for two new tests. *Br. J. Psychol.* 101, 563–578. doi: 10.1348/000712609X474370

Austin, E. J., Parker, J. D. A., Petrides, K. V., and Saklofske, D. H. (2008). "Emotional intelligence," in *He SAGE Handbook of Personality Theory and Assessment*, eds G. J. Boyle, G. Matthews, and D. H. Saklofske (London: SAGE), 576–596.

Baker, F. B. (1985). *The Basics of Item Response Theory*. Portsmouth, NH: Heineman.

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (London: Addison-Wesley), 397–479.

Brackett, M. A., Rivers, S. E., Shiffman, S., Lerner, N., and Salovey, P. (2006). Relating emotional abilities to social functioning: a comparison of self-report and performance measures of emotional intelligence. *J. Pers. Soc. Psychol.* 91, 780–795. doi: 10.1037/0022-3514.91.4.780

Brackett, M. A., Warner, R. M., and Bosco, J. S. (2005). Emotional intelligence and relationship quality among couples. *Pers. Relatsh.* 12, 197–212. doi: 10.1111/j.1350-4126.2005.00111.x

Cho, S., Drasgow, F., and Cao, M. (2015). An investigation of emotional intelligence measures using item response theory. *Psychol. Assess.* 27, 1241–1252. doi: 10.1037/pas0000132

Ciarrochi, J., Deane, F. P., and Anderson, S. (2002). Emotional intelligence moderates the relationship between stress and mental health. *Pers. Individ. Dif.* 32, 197–209. doi: 10.1016/S0191-8869(01)00012-5

Côté, S. (2014). Emotional intelligence in organizations. *Annu. Rev. Organ. Psychol. Organ. Behav.* 1, 459–488. doi: 10.1146/annurev-orgpsych-031413-091233

Côté, S., DeCelles, K. A., McCarthy, J. M., Van Kleef, G. A., and Hideg, I. (2011). The jekyll and hyde of emotional intelligence: emotion-regulation knowledge facilitates both prosocial and interpersonally deviant behavior. *Psychol. Sci.* 22, 1073–1080. doi: 10.1177/0956797611416251

Dimitrijević, A. A., Marjanović, Z. J., and Dimitrijević, A. (2018). Whichever intelligence makes you happy: the role of academic, emotional, and practical abilities in predicting psychological well-being. *Pers. Individ. Dif.* 132, 6–13. doi: 10.1016/j.paid.2018.05.010

Ferguson, F. J., and Austin, E. J. (2010). Associations of trait and ability emotional intelligence with performance on theory of mind tasks in an adult sample. *Pers. Individ. Dif.* 49, 414–418. doi: 10.1016/j.paid.2010.04.009

Gallagher, E. N., and Vella-Brodrick, D. A. (2008). Social support and emotional intelligence as predictors of subjective well-being. *Pers. Individ. Dif.* 44, 1551–1561. doi: 10.1016/j.paid.2008.01.011

Greenidge, D., Devonish, D., and Alleyne, P. (2014). The relationship between ability-based emotional intelligence and contextual performance and counterproductive work behaviors: a test of the mediating effects of job satisfaction. *Hum. Perform.* 27, 225–242. doi: 10.1080/08959285.2014.913591

Joseph, D. L., and Newman, D. A. (2010). Emotional intelligence: an integrative meta-analysis and cascading model. *J. Appl. Psychol.* 95, 54–78. doi: 10.1037/a0017286

Kafetsios, K., and Zampetakis, L. A. (2008). Emotional intelligence and job satisfaction: testing the mediatory role of positive and negative affect at work. *Pers. Individ. Dif.* 44, 712–722. doi: 10.1016/j.paid.2007.10.004

Karim, J. (2010). An item response theory analysis of wong and law emotional intelligence scale. *Procedia Soc. Behav. Sci.* 2, 4038–4047. doi: 10.1016/j.sbspro.2010.03.637

Leavitt, K., Fong, C. T., and Greenwald, A. G. (2011). Asking about well-being gets you half an answer: intra-individual processes of implicit and explicit job attitudes. *J. Organ. Behav.* 32, 672–687. doi: 10.1002/job.746

Li, H., Zhang, Y., and Li, F. (2017). Psychometric properties of the multi-affect indicator in a Chinese worker sample. *Psychol. Rep.* 120, 179–188. doi: 10.1177/0033294116676464

Libbrecht, N., and Lievens, F. (2012). Validity evidence for the situational judgment test paradigm in emotional intelligence measurement. *Int. J. Psychol.* 47, 438–447. doi: 10.1080/00207594.2012.682063

Libbrecht, N., Lievens, F., Carette, B., and Côté, S. (2014). Emotional intelligence predicts success in medical school. *Emotion* 14, 64–73. doi: 10.1037/a0034392

MacCann, C., and Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion* 8, 540–551. doi: 10.1037/a0012746

Mayer, J. D., and Salovey, P. (1997). "What is emotional intelligence?," in *Emotional Development and Emotional Intelligence: Educations Implications*, eds P. Salovey and D. Sluyter (New York, NY: Harper Collins), 3–31.

Mayer, J. D., Salovey, P., and Caruso, D. R. (2002). *Mayer-Salovey-Caruso Emotional Intelligence Test MSCEIT User's Manual*. New York, NY: Multi-Health Systems Inc.

Mayer, J. D., Salovey, P., Caruso, D. R., and Sitarenios, G. (2001). Emotional intelligence as a standard intelligence. *Emotion* 1, 232–242. doi: 10.1037//1528-3542.1.3.232-242

Meisler, G. (2014). Exploring emotional intelligence, political skill, and job satisfaction. *Employee Relat.* 36, 280–293. doi: 10.1108/ER-02-2013-0021

Miao, C., Humphrey, R. H., and Qian, S. (2017). A meta-analysis of emotional intelligence and work attitudes. *J. Occup. Organ. Psychol.* 90, 177–202. doi: 10.1111/joop.12167

Miao, C., Humphrey, R. H., and Qian, S. (2018). A cross-cultural meta-analysis of how leader emotional intelligence influences subordinate task performance and organizational citizenship behavior. *J. World Bus.* 53, 463–474. doi: 10.1016/j.jwb.2018.01.003

Moon, S. (2011). *East meets West: The Cultural-Relativity of Emotional Intelligence*. Doctoral dissertation, University of Toronto, Toronto

Mulki, J. P., Jaramillo, F., Goad, E. A., and Pesquera, M. R. (2015). Regulation of emotions, interpersonal conflict, and job performance for salespeople. *J. Bus. Res.* 68, 623–630. doi: 10.1016/j.jbusres.2014.08.009

Newman, D. A., Joseph, D. L., and MacCann, C. (2010). Emotional intelligence and job performance: the importance of emotion regulation and emotional labor context. *Ind. Organ. Psychol.* 3, 159–164. doi: 10.1111/j.1754-9434.2010.01218.x

O'Boyle, E. H., Humphrey, R. H., Pollack, J. M., Hawver, T. H., and Story, P. A. (2011). The relation between emotional intelligence and job performance: a meta-analysis. *J. Organ. Behav.* 32, 788–818. doi: 10.1002/job.714

Oginska-Bulik, N. (2005). Emotional intelligence in the workplace: exploring its effects on occupational stress and health outcomes in human service workers. *Int. J. Occup. Med. Environ. Health* 18, 167–175.

Ouyang, Z., Sang, J., Li, P., and Peng, J. (2015). Organizational justice and job insecurity as mediators of the effect of emotional intelligence on job satisfaction: a study from china. *Pers. Individ. Dif.* 76, 147–152. doi: 10.1016/j.paid.2014.12.004

Parke, M. R., Seo, M. G., and Sherf, E. N. (2015). Regulating and facilitating: the role of emotional intelligence in maintaining and using positive affect for creativity. *J. Appl. Psychol.* 100, 917–934. doi: 10.1037/a0038452

Rizopoulos, D. (2006). ltm: an R package for latent variable modeling and item response theory analyses. *J. Stat. Softw.* 17, 1–25. doi: 10.18637/jss.v017.i05

Roberts, R. D., Schulze, R., O'Brien, K., MacCann, C., Reid, J., and Maul, A. (2006). Exploring the validity of the Mayer–Salovey–Caruso emotional intelligence test (MSCEIT) with established emotions measures. *Emotion* 6, 663–669. doi: 10.1037/1528-3542.6.4.663

Roseman, I. J. (2001). "A model of appraisal in the emotion system: integrating theory, research, and applications," in *Appraisal Processes in Emotion: Theory, Methods, Research*, eds K. R. Scherer and A. Schorr (New York: Oxford University Press), 68–91.

Schriesheim, C., and Tsui, A. S. (1980). "Development and validation of a short satisfaction instrument for use in survey feedback interventions," in *Paper presented at the Western Academy of Management Meeting*, London.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., et al. (1998). Development and validation of a measure of emotional intelligence. *Pers. Individ. Dif.* 25, 167–177. doi: 10.1016/S0191-8869(98)00001-4

Sheldon, O. J., Ames, D. R., and Dunning, D. (2014). Emotionally unskilled, unaware, and uninterested in learning more: reactions to feedback about deficits in emotional intelligence. *J. Appl. Psychol.* 99, 125–137. doi: 10.1037/a0034138

Sy, T., Tram, S., and O'Hara, L. A. (2006). Relationship of employee and manager emotional intelligence to job satisfaction and performance. *J. Vocat. Behav.* 68, 461–473. doi: 10.1016/j.jvb.2005.10.003

Van Rooy, D. L., and Viswesvaran, C. (2004). Emotional intelligence: a meta-analytic investigation of predictive validity and nomological net. *J. Vocat. Behav.* 65, 71–95. doi: 10.1016/S0001-8791(03)00076-9

Vratskikh, I., Al-Lozi, M., and Maqableh, M. (2016). The impact of emotional intelligence on job performance via the mediating role of job satisfaction. *Int. J. Bus. Manag.* 11, 69–91. doi: 10.5539/ijbm.v11n2p69

Wang, L., and Lin, W. (2011). Wording effects and the dimensionality of the general health questionnaire (GHQ-12). *Pers. Individ. Dif.* 50, 1056–1061. doi: 10.1016/j.paid.2011.01.024

Warr, P., Bindl, U. K., Parker, S. K., and Inceoglu, I. (2014). Four-quadrant investigation of job-related affects and behaviours. *Eur. J. Work Organ. Psychol.* 23, 342–363. doi: 10.1080/1359432X.2012.744449

Warr, P., and Parker, S. (2010). *IWP Multi-Affect Indicator*. Sheffield: University of Sheffield.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 14 June 2019
doi: 10.3389/fpsyg.2019.01135

Development and Validation of Verbal Emotion Vignettes in Portuguese, English, and German

Tanja S. H. Wingenbach, Leticia Y. Morello, Ana L. Hack and Paulo S. Boggio*

Social and Cognitive Neuroscience Laboratory, Centre for Biological and Health Sciences, Mackenzie Presbyterian University, São Paulo, Brazil

Everyday human social interaction involves sharing experiences verbally and these experiences often include emotional content. Providing this context generally leads to the experience of emotions in the conversation partner. However, most emotion elicitation stimulus sets are based on images or film-sequences providing visual and/or auditory emotion cues. To assimilate what occurs within social interactions, the current study aimed at creating and validating verbal emotion vignettes as stimulus set to elicit emotions (anger, disgust, fear, sadness, happiness, gratitude, guilt, and neutral). Participants had to mentally immerse themselves in 40 vignettes and state which emotion they experienced next to the intensity of this emotion. The vignettes were validated on a large sample of native Portuguese-speakers (N = 229), but also on native English-speaking (N = 59), and native German-speaking (N = 50) samples to maximise applicability of the vignettes. Hierarchical cluster analyses showed that the vignettes mapped clearly on their target emotion categories in all three languages. The final stimulus sets each include 4 vignettes per emotion category plus 1 additional vignette per emotion category which can be used for task familiarisation procedures within research. The high agreement rates on the experienced emotion in combination with the medium to large intensity ratings in all three languages suggest that the stimulus sets are suitable for application in emotion research (e.g., emotion recognition or emotion elicitation).

Keywords: emotion vignettes, emotion, German, Portuguese, English

OPEN ACCESS

Edited by:
Elisa Pedroli,
Istituto Auxologico Italiano (IRCCS),
Italy

Reviewed by:
Fernando Marmolejo-Ramos,
The University of Adelaide, Australia
Cindy Harmon-Jones,
University of New South Wales,
Australia

****Correspondence:***
Tanja S. H. Wingenbach
tanja.wingenbach@bath.edu

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *28 January 2019*
Accepted: *30 April 2019*
Published: *14 June 2019*

Citation:
Wingenbach TSH, Morello LY, Hack AL and Boggio PS (2019) Development and Validation of Verbal Emotion Vignettes in Portuguese, English, and German. Front. Psychol. 10:1135. doi: 10.3389/fpsyg.2019.01135

INTRODUCTION

The everyday life of humans involves many social interactions which are rarely free of emotional content. When we interact with each other, we tell stories about experiences including emotional states, and use facial expressions to communicate about our emotional states in addition to varying intonation and speed of our speech. Thus, a multitude of stimulus sets providing sensory cues exist for investigation of related research questions, e.g., stimulus sets of facial emotion (literature review by Ekman and Friesen, 1976; Langner et al., 2010; Krumhuber et al., 2013; Wingenbach et al., 2016) and vocalisations (Belin et al., 2008) but also including multiple modalities (Bänziger et al., 2009, 2012; Hawk et al., 2009; Dyck, 2012). Such stimulus sets are useful when investigating participants' processing of other's emotions based on sensory information and are generally stripped of contextual information.

Stimuli including contextual information are more likely to elicit an emotion in the observer or listener. Stimulus sets have accordingly been developed with the purpose to elicit emotions. A widely used stimulus set is the International Affect Picture Set (IAPS; Lang et al., 1997) which includes thousands of images depicting emotional scenes validated to elicit affect ranging in valence from negative to positive (Ito et al., 1998). There are also dynamic stimulus sets that can elicit affect, e.g., a film-based stimulus set containing 20 stimuli of positive vs. negative social interactions (Carvalho et al., 2012). Whereas these stimulus sets range on the valence dimension, there are also stimulus sets that aim at the elicitation of specific emotions, e.g., emotion eliciting film sequences (McHugo et al., 1982; Philippot, 1993; Gross and Levenson, 1995; Schaefer et al., 2010).

Emotion-specific stimulus sets often include the six emotion categories which are agreed upon by most researchers to represent so called basic emotions (Ekman et al., 1969; Ekman and Cordaro, 2011), but see also (Ortony and Turner, 1990). These emotions are anger, disgust, sadness, fear, happiness, and surprise. Because these emotions are considered universal, i.e., culturally independent, their inclusion in stimulus sets is often standard. However, many more emotions exist and are often called complex emotions, since they include a greater cognitive component than basic emotions. Examples of complex emotions are gratitude and guilt. To be able to experience gratitude, it is necessary to evaluate an action by someone else as beneficial to oneself and costly to the other person at the same time (McCullough et al., 2008). It is this saccade of appraisals that makes gratitude a complex emotion. The same applies to guilt. Here, an action carried out by oneself might have been beneficial to oneself but included negative aspects for another person (Tracy and Robins, 2006). Guilt as well as gratitude are emotions that emerge in interpersonal contexts and are thus of great interest to social psychology research. The authors are unaware of a stimulus set suitable for elicitation of emotions including these two complex emotions next to basic emotions. It is possible that it is difficult to induce guilt and gratitude with images whether static or dynamic and that therefore the focus is on basic emotions within such stimulus sets.

As opposed to watching films or images, reporting about experiences in conversations within social interactions includes verbal descriptions of scenarios. A semantic understanding by the listener is required as well as abilities of perspective taking to understand the emotional experience of the narrator and to experience their emotions. Verbal vignettes depicting brief situations of emotional content are a useful research tool incorporating these aspects. The "Geneva Emotion Knowledge test – Blends" includes 28 verbal vignettes each portraying two out of 16 target emotions (pride, joy, happiness, pleasure, interest, anxiety, sadness, irritation, fear, disgust, anger, guilt, shame, contempt, jealousy, and surprise). These vignettes can be used to measure emotion understanding (Schlegel and Scherer, 2017). When participants are instructed to mentally immerse themselves in the described scenarios, it is possible to elicit emotion experience. For example, a published study taking this approach included one verbal vignette depicting five emotions (anger, sadness, jealousy, embarrassment, and anxiety) (Vine et al., 2018). Whereas the individual vignettes used by Schlegel and Scherer (2017) and Vine et al. (2018) included several target emotions, it is also possible to target specific emotions one at a time within individual vignettes.

Verbal vignettes describing situations of one target emotion each (anger, sadness, and fear) were created by MacCann and Roberts (2008) and Hareli et al. (2011), the latter included vignettes depicting guilt. The International Survey on Emotion Antecedents and Reactions (Scherer and Wallbott) is a database of situations described by almost 3000 participants that elicited a specific emotion in them (joy, fear, anger, sadness, disgust, shame, and guilt). Whereas guilt as a target emotion is sometimes included alongside other emotions, vignettes targeting gratitude are generally not included. However, there is published research which focussed on gratitude itself. For example, a study included three gratitude vignettes although two of these vignettes described the same situation but was varied in the intensity of the received benefit (Wood et al., 2008) and another study included 12 gratitude vignettes (Lane and Anderson, 1976). The authors are unaware of a vignette stimulus set including gratitude and guilt next to basic emotions.

The current research aimed at developing and validating verbal emotion vignettes of seven different emotion categories alongside neutral vignettes. To assure that the vignettes can induce emotions, high agreement rates on the experienced emotions, and intensity ratings were necessary. Thus, agreement rates and intensity rates were calculated per vignette. It was required for each individual vignette to distinctively map onto one emotion category based on the agreement rates, which was addressed with hierarchical clustering. Based on the agreement rates, hit rates (raw and unbiased), and intensity rates were calculated for each emotion category for comparison to published instruments. To increase the benefit of the emotion vignettes to the research community, the vignettes were created, and validated in three languages (Portuguese, English, and German).

MATERIALS AND METHODS

Stimuli Creation

Verbal vignettes were created written from a first-person perspective to facilitate for the reader to imagine the situation described in the vignettes. The vignettes were each written with a similar length of ∼3 lines. It was aimed to describe scenarios that would clearly map onto one distinct emotion category. Initially, 10 vignettes were created per emotion category (anger, disgust, fear, sadness, guilt, happiness, and gratitude) and also for neutral scenarios. Several pilot studies were conducted on psychology student samples. Each pilot study led to adjustments of the wording of the vignettes and clarification of the task instructions with the aim to increase recognition rates of the individual vignettes. Every vignette with a recognition rate of the target emotion <80% was re-written to be more distinct.

Eventually, 5 vignettes per emotion category with recognition rates of > = 80% were selected to be included in the validation study (presented in the results section of the current manuscript). The vignettes with the highest recognition rates were selected, as the aim for the vignettes was to have as little ambiguity as possible. All 40 vignettes in each of the three languages can be found in the **Supplementary Material** but example vignettes (one for each emotion category) are provided in the following:

> Anger: "I was eating cake at home with my sister when her boyfriend arrived. He glanced at the cake and said she should stop eating because she was getting too fat and he wouldn't date her anymore if she continued like that."
> Disgust: "On my way home, I saw a dead rat on the sidewalk. When I got closer I noticed its belly was open, decomposing, with tons of white maggots crawling inside it, and some coming out of its mouth."
> Fear: "It was late one night, and I was in a deserted plaza with some friends. We were laughing and walking in the direction of the car when my friend was struck in the back. We all froze when we saw two men pointing guns at us."
> Sadness: "When me and my sister were younger, we became orphans. We ended up being sent to different homes. I remember this day, because my sister cried a lot and held me tight. I didn't understand why I couldn't stay with her."
> Guilt: "When I ended my relationship, I shared intimate photos of my ex-girlfriend with a group of friends. These pictures were leaked to the internet, and afterward I found out she had been fired from her job for getting a bad reputation. I should never have done that."
> Neutral: "I left college at noon and went to the parking lot to pick up my car and leave. On the way, there was a restaurant and I had lunch there before heading on. I got on my way and home at two o'clock."
> Happiness: "I went to see a show of a band I've been a fan of since I was a teenager. During the show, the vocalist saw my poster, walked toward me smiling, and reached out to me while singing my favourite song."
> Gratitude: "Late one night, I slept on the last bus and only woke up at the final bus stop. My cell phone battery was dead and, hearing my story, a station worker let me borrow his phone to call someone."

Participants

Portuguese-speaking participants were recruited from the Mackenzie Presbyterian University student population through social media. Data was collected from 301 participants. A control measure was inserted in the online assessment to identify participants who did not pay attention to their answering. After exclusion of these individuals, the final sample size included in the analyses was N = 229 [202 females, 27 males; M(age) = 20.7 years, SD = 4.7]. English-speaking participants [N = 59, 30 females, 29 males; M(age) = 34.5 years, SD = 10.9]

TABLE 1 | Agreement rates and intensity rates in percentages for each vignette in English, Portuguese, and German.

		Anger				**Disgust**				**Fear**				**Sadness**		
	#	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***	**#**	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***	**#**	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***	**#**	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***
A	**1**	71 (46)	95 (21)	92 (27)	**6**	83 (38)	87 (34)	84 (37)	**11**	90 (31)	97 (16)	96 (20)	**16**	81 (39)	89 (31)	78 (42)
I		71 (18)	80 (22)	68 (22)		79 (23)	82 (20)	72 (22)		87 (20)	94 (13)	84 (23)		81 (24)	88 (20)	83 (25)
A	**2**	71 (46)	91 (29)	88 (38)	**7**	80 (41)	87 (34)	78 (42)	**12**	86 (35)	97 (17)	84 (37)	**17**	80 (41)	88 (33)	78 (42)
I		81 (20)	87 (18)	75 (24)		76 (22)	73 (22)	65 (24)		86 (24)	92 (16)	83 (22)		82 (21)	85 (21)	79 (25)
A	**3**	68 (47)	86 (35)	92 (27)	**8**	81 (39)	92 (28)	64 (49)	**13**	86 (35)	94 (23)	88 (33)	**18**	76 (43)	86 (34)	70 (46)
I		74 (21)	90 (15)	71 (20)		57 (25)	79 (24)	58 (24)		66 (25)	80 (21)	70 (22)		85 (20)	91 (17)	81 (22)
A	**4**	64 (48)	72 (45)	80 (40)	**9**	81 (39)	94 (23)	88 (33)	**14**	81 (39)	91 (28)	88 (33)	**19**	75 (44)	93 (26)	70 (46)
I		59 (25)	73 (24)	65 (20)		81 (22)	87 (20)	79 (19)		84 (23)	95 (11)	82 (23)		82 (23)	92 (14)	83 (20)
A	**5**	47 (50)	81 (39)	68 (47)	**10**	73 (45)	83 (38)	56 (50)	**15**	66 (48)	89 (31)	80 (40)	**20**	69 (46)	77 (42)	50 (51)
I		60 (26)	76 (22)	52 (20)		70 (27)	75 (24)	56 (27)		50 (21)	74 (25)	59 (27)		69 (25)	86 (19)	76 (19)
		Happiness				**Gratitude**				**Guilt**				**Neutral**		
	#	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***	**#**	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***	**#**	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***	**#**	**ENG** ***M (SD)***	**POR** ***M (SD)***	**GER** ***M (SD)***
A	**21**	83 (38)	93 (45)	84 (37)	**26**	83 (38)	91 (28)	84 (37)	**31**	76 (43)	81 (40)	84 (37)	**36**	85 (36)	92 (27)	84 (37)
I		72 (24)	86 (19)	79 (22)		62 (25)	88 (17)	74 (20)		82 (22)	90 (16)	79 (22)		49 (32)	69 (37)	51 (36)
A	**22**	81 (39)	90 (31)	98 (14)	**27**	83 (38)	92 (28)	84 (37)	**32**	75 (44)	72 (45)	86 (35)	**37**	80 (41)	72 (45)	78 (42)
I		78 (21)	90 (16)	79 (19)		68 (22)	87 (16)	70 (19)		78 (23)	93 (12)	84 (20)		56 (33)	66 (35)	40 (35)
A	**23**	80 (41)	91 (28)	90 (30)	**28**	80 (41)	94 (24)	84 (37)	**33**	73 (45)	76 (43)	78 (42)	**38**	75 (44)	90 (31)	90 (30)
I		70 (26)	86 (17)	80 (20)		70 (22)	88 (17)	77 (17)		78 (32)	92 (15)	70 (27)		54 (35)	65 (36)	42 (33)
A	**24**	90 (31)	79 (41)	76 (43)	**29**	76 (43)	90 (31)	84 (37)	**34**	68 (47)	77 (42)	82 (39)	**39**	80 (41)	65 (48)	88 (33)
I		70 (24)	86 (18)	76 (19)		68 (22)	81 (20)	76 (18)		89 (19)	89 (16)	87 (17)		44 (30)	65 (34)	41 (37)
A	**25**	73 (45)	87 (34)	76 (43)	**30**	69 (46)	78 (42)	80 (40)	**35**	39 (49)	68 (47)	66 (48)	**40**	75 (44)	88 (33)	80 (40)
I		66 (24)	81 (21)	70 (25)		72 (23)	86 (16)	73 (18)		87 (15)	95 (11)	88 (18)		55 (33)	66 (37)	43 (34)

EMO, emotion; ENG, English; POR, Portuguese; GER, German; M, mean; SD, standard deviation; A, recognition rate; I, intensity rate.

were recruited through social media from the general population. English as mother tongue was required for participation in the study. German-speaking participants [*N* = 50, 28 females, 22 males, *M*(age) = 37.4 years, *SD* = 11.7] were recruited from the general population through social media and German as mother tongue was a requirement for study participation. No participants were excluded from the English-speaking and German-speaking samples for analyses.

Procedure

Ethical approval of the study was provided by the Mackenzie Presbyterian University Ethics Committee. Participants accessed the vignettes through a Google Forms survey and written informed consent for participation was obtained within the survey. Participants were instructed to participate from a place without distractions, to answer on their own, and not to engage in any other activity while completing the study. The instruction for each vignette was for the reader to imagine to be the person depicted in the scenario and immerse themselves in the scenario. Participants then had to choose one emotion category from a list of provided labels (one for each of the 8 emotion categories) to state what they were feeling while they imaged to experience the situation depicted in the vignette. Next, participants had to rate the intensity of the chosen emotion for the respective vignette on a 10-point Likert-scale ranging from 0 (=very low) to 9 (=very high). Completing the study took approximately 25 min. Portuguese-speaking participants were granted course credit for participation. English-speaking and German-speaking participants were not compensated for participation as required by Brazilian law.

Statistical Methods

Data files (one for each language) were created including participants' responses to each vignette. The responses to the first question (emotion label attributions) for each vignette were

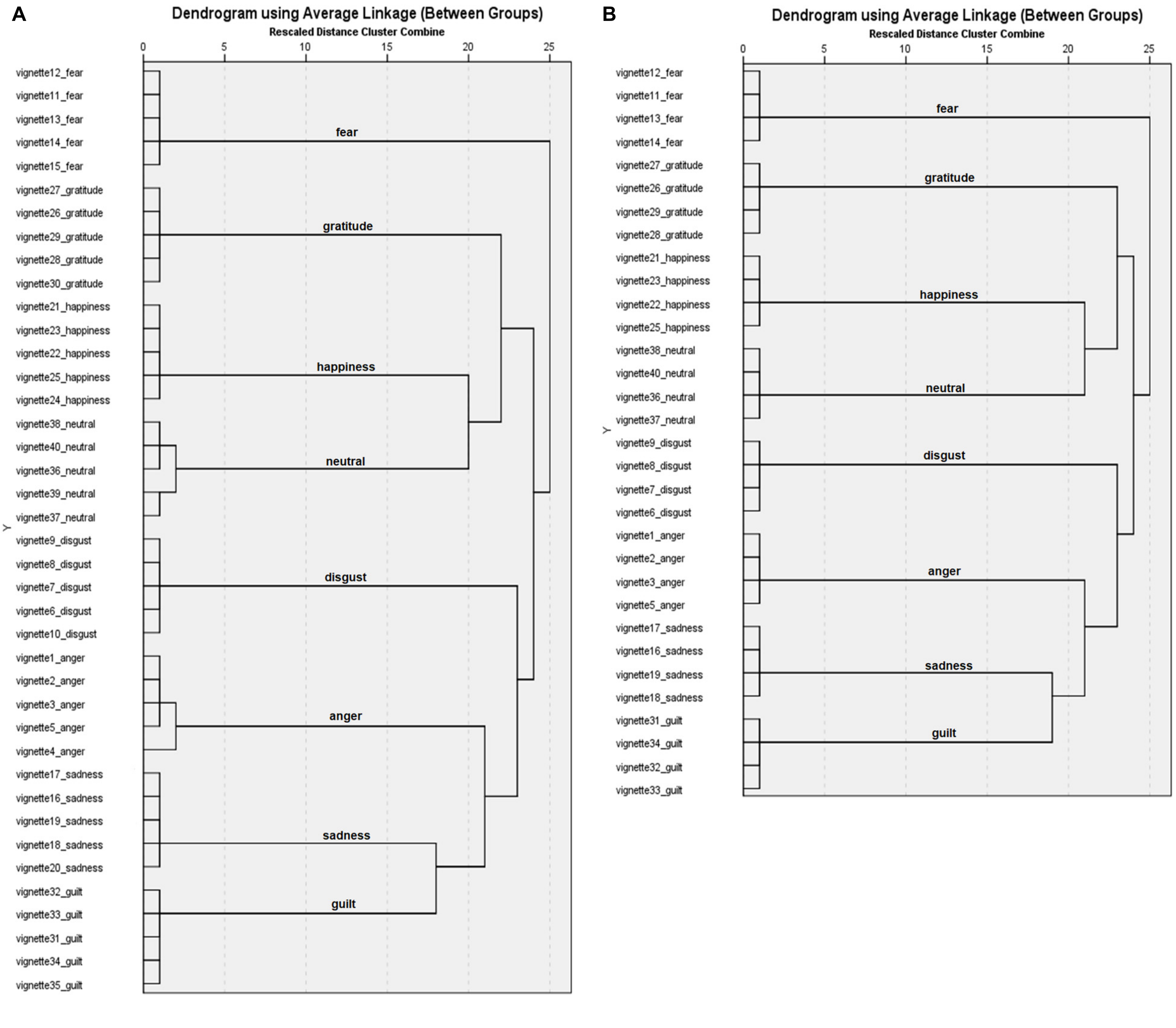


FIGURE 1 | Dendrograms for the Portuguese vignettes with **(A)** 40 vignettes and **(B)** 32 vignettes.

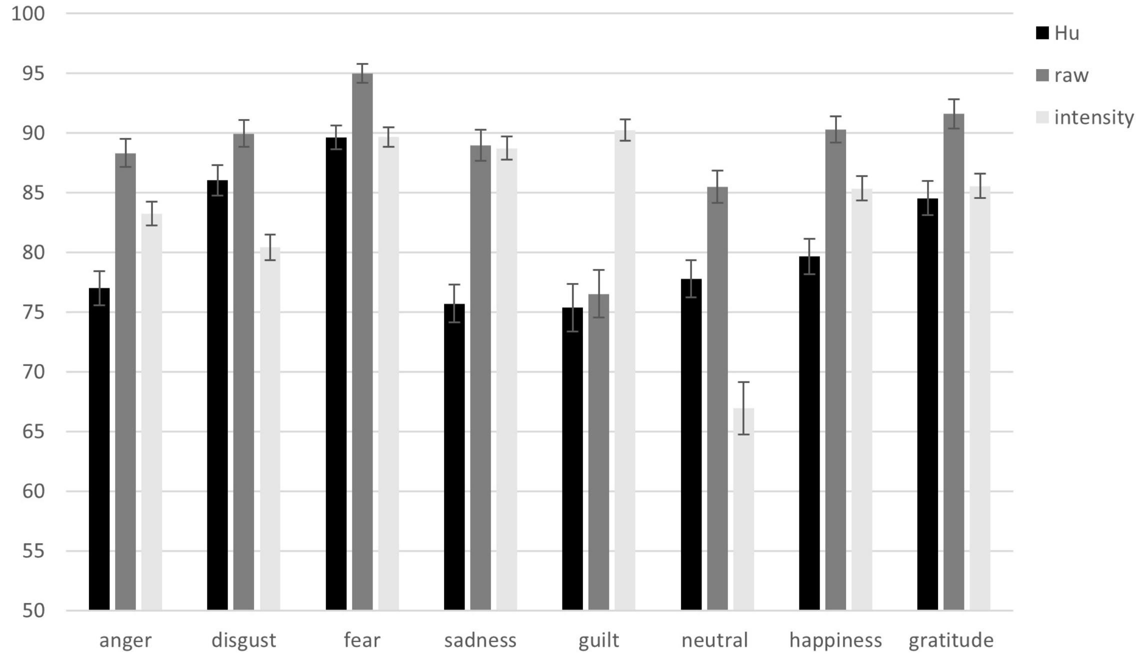


FIGURE 2 | Unbiased hit rates (*Hu*), raw hit rates, and intensity rates from the Portuguese vignettes validation per emotion category. Error bars represent standard errors of the means.

transformed to reflect target emotion attributions by assigning ones and non-target attributions by assigning zeros to be able to calculate raw hit rates per vignette (separately for each language). That is, for each vignette, the number of attributions of the target emotion across participants was summed, divided by the respective *N*, and multiplicated by 100 (i.e., rule of three, to represent percentages for ease of interpretation). Likewise, mean intensity rates (in %) per vignette were calculated (only considering classifications of the target emotion to the individual vignettes) by applying the rule of three, i.e., the intensity ratings of all participants were averaged per vignette, divided by 9, and multiplicated by 100.

Statistical analyses were conducted using the software SPSS (version 24; IBM Corp, 2016). A hierarchical cluster analysis with average linkage between groups and squared Euclidian distance was conducted (separately for each language) including all 40 vignettes to test whether the individual vignettes clearly mapped onto one emotion category as intended based on the sum of emotion label attributions per category (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude). Vignettes that did not clearly map onto their target emotion category were eliminated and the hierarchical cluster analysis was conducted again only including the remaining vignettes.

Afterward, raw hit rates per emotion category were calculated (separately for each language) by averaging the raw hit rates (in %) of the four vignettes per emotion category to be included in the final stimuli sets as identified by the cluster analyses.

As a measure of distinctiveness, unbiased hit rates (Hu; Wagner, 1993) were calculated for each emotion category (separately for each language). *Hu* takes response biases into consideration by which the raw hit rates are corrected. The formula is $Hu = a^2/(a + b + c)^*(a + d + e)$ where *a* represents the target emotion, *b* and *c* represent the misattributions of another emotion to the presented target emotion, and *d* and *e* represent the misattributions of the target emotion to other emotion categories. The resulting *Hu* rates represent percentages.

Intensity rates were calculated per emotion category (separately for each language) by averaging the intensity rates (in %) of the four vignettes per emotion category as identified by the cluster analyses to be included in the final stimuli sets.

RESULTS

Portuguese Vignettes

Table 1 displays the *M*s and *SD*s of the raw hit rates and intensity rates for the individual vignettes.

Cluster Analyses

Results (**Figure 1A**) from the hierarchical cluster analysis showed that for 6 emotion categories (disgust, fear, sadness, happiness, gratitude, and guilt) all 5 emotion vignettes for the target emotion categories were clustered together on the first cluster level. For 2 categories (neutral and anger), clusters emerged on first, and second level. After eliminating the vignette with the lowest recognition rates for each of the 8 emotion categories, cluster analysis including 32 vignettes showed 8 clusters including 4 vignettes each at the first level (**Figure 1B**). The single solution of 8 clusters also grouped all vignettes according to their target emotion. All following results are based on the 4 identified vignettes per emotion category.

TABLE 2 | Confusions between the emotion categories in percentages from all three studies.

		Responses							
		Anger	**Disgust**	**Fear**	**Sadness**	**Happiness**	**Neutral**	**Gratitude**	**Guilt**
Portuguese									
Target emotion	Anger	88	3	1	6	0	2	0	0
	Disgust	2	90	1	2	1	4	0	0
	Fear	2	0	95	2	0	0	0	0
	Sadness	3	0	4	89	1	2	0	1
	Happiness	0	0	0	0	90	3	6	0
	Neutral	0	0	1	0	9	86	4	0
	Gratitude	0	1	0	0	5	2	92	0
	Guilt	9	2	1	11	0	1	0	77
English									
Target emotion	Anger	69	6	10	3	3	4	1	3
	Disgust	2	81	2	1	3	6	3	2
	Fear	2	0	86	1	2	3	2	3
	Sadness	2	3	5	78	3	3	3	4
	Happiness	1	2	2	0	83	3	7	1
	Neutral	1	1	1	1	9	80	5	3
	Gratitude	1	3	3	1	4	4	81	4
	Guilt	3	2	4	10	3	2	3	73
German									
Target emotion	Anger	88	0	5	1	0	4	0	1
	Disgust	1	78	1	3	4	11	0	0
	Fear	6	0	89	0	0	3	0	0
	Sadness	9	0	8	74	0	4	0	4
	Happiness	0	0	0	1	90	3	6	0
	Neutral	0	0	0	0	10	86	2	1
	Gratitude	0	0	0	1	11	3	84	0
	Guilt	4	0	1	6	1	4	0	83

All values are based on four vignettes per target emotion category. The percentages in the diagonal line represent the raw hit rates; all percentages below and above the diagonal represent the percentages of confusions.

Raw Hit Rates per Emotion Category

Raw hit rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 2**.

Hu Rates per Emotion Category

Hu rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 2**. The confusions between emotion categories underlying the *Hu* rates are presented in **Table 2**.

Intensity Rates per Emotion Category

Intensity rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 2**.

English Vignettes

Table 1 displays the *M*s and *SD*s of the raw hit rates and intensity rates for the individual vignettes.

Cluster Analyses

Results (**Figure 3A**) from the hierarchical cluster analysis showed that for 5 emotion categories (disgust, sadness, gratitude, happiness, and neutral) all 5 emotion vignettes for the target emotion categories were clustered together on the first cluster level. For 3 categories (fear, anger, and guilt), 4 vignettes were categorised as belonging together on the first cluster level and 1 vignette was clustered to the target category on higher levels (level 2 and level 5). After eliminating the vignette with the lowest recognition rates for each of the 8 emotion categories, cluster analysis including 32 vignettes showed 8 clusters including 4 vignettes each at the first cluster level (**Figure 3B**). The single solution of 8 clusters also grouped all vignettes according to their target emotion. All following results are based on the 4 identified vignettes per emotion category.

Raw Hit Rates per Emotion Category

Raw hit rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 4**.

Hu Rates per Emotion Category

Hu rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude)

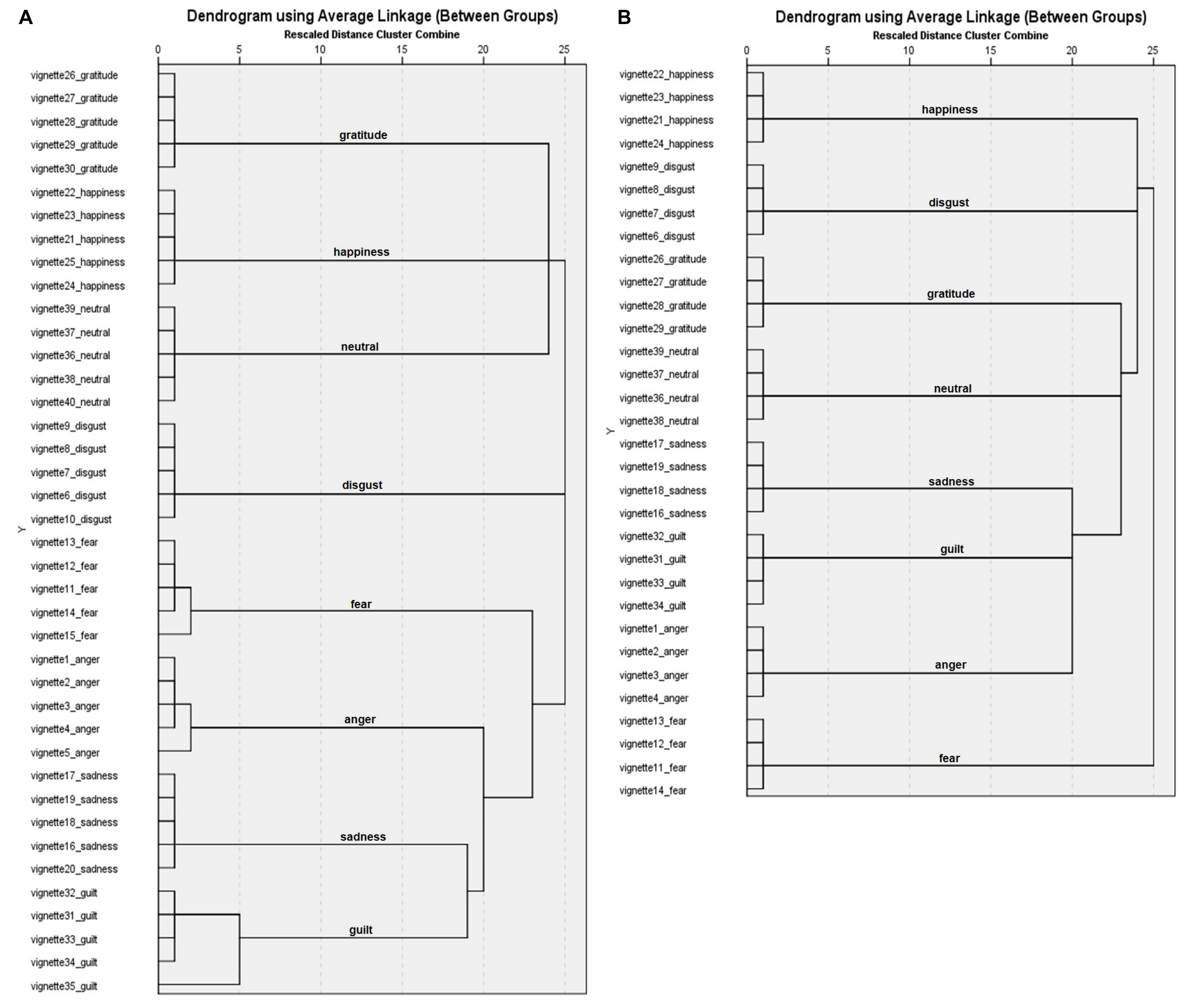


FIGURE 3 | Dendrograms for the English vignettes with **(A)** 40 vignettes and **(B)** 32 vignettes.

are presented in **Figure 4**. The confusions between emotion categories underlying the *Hu* rates are presented in **Table 2**.

Intensity Rates per Emotion Category

Intensity rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 4**.

German Vignettes

Table 1 displays the *M*s and *SD*s of the raw hit rates and intensity rates for the individual vignettes.

Cluster Analyses

Results (**Figure 5A**) from the cluster analysis showed that for 5 emotion categories (fear, gratitude, happiness, guilt, and neutral) all 5 emotion vignettes for the target emotion categories were clustered together. For 2 categories (sadness and anger), 4 stories were categorised as belonging together on the first cluster level and one story was clustered to the target emotion at a higher level (level 2 and level 3). For the category of disgust, 3 clusters emerged ranging from level 1 to 3. After eliminating the vignette with the lowest recognition rates for each of the 8 emotion categories, cluster analysis including 32 vignettes showed 7 clusters including 4 vignettes each at the first level and there was a second cluster between disgust vignettes at the second level (**Figure 5B**). The single solution of 8 clusters grouped all vignettes according to their target emotion including disgust. All following results are based on the 4 identified vignettes per emotion category.

Raw Hit Rates per Emotion Category

Raw hit rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 6**.

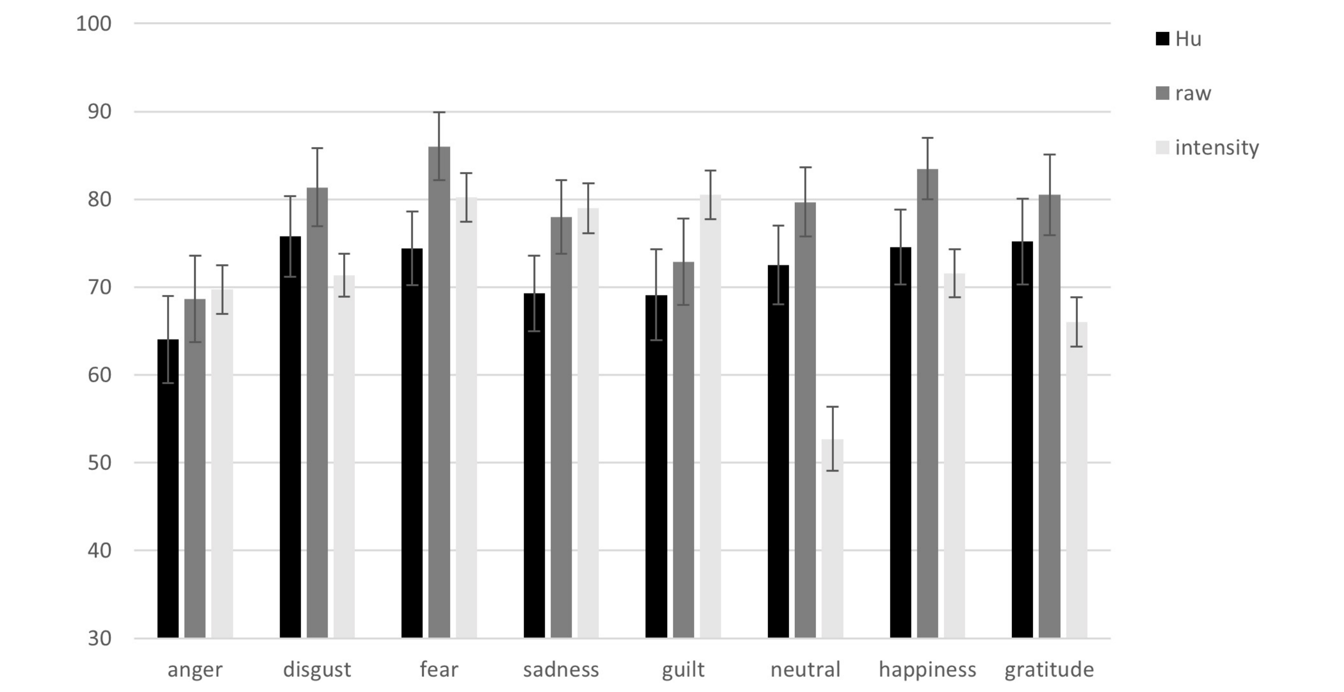


FIGURE 4 | Unbiased hit rates (*Hu*), raw hit rates, and intensity rates from the English vignettes validation per emotion category. Error bars represent standard errors of the means.

Hu Rates per Emotion Category

Hu rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 6**. The confusions between emotion categories underlying the *Hu* rates are presented in **Table 2**.

Intensity Rates per Emotion Category

Intensity rates (*M*s and *SE*s) for the emotion categories (anger, disgust, fear, sadness, guilt, neutral, happiness, and gratitude) are presented in **Figure 6**.

DISCUSSION

The current research aimed at developing and validating verbal vignettes portraying short scenarios related to the specific emotions of anger, disgust, sadness, fear, happiness, gratitude, guilt, and neutral. Results showed that the individual emotion vignettes included in the final stimulus sets clearly mapped onto distinct emotion categories for each of the three languages. Results further showed high intensity rates for the self-reported experience of emotions while participants immersed themselves in the scenarios depicted in the vignettes. The vignettes can thus be considered successfully validated making them applicable within emotion research, e.g., emotion recognition and emotion elicitation.

When including five vignettes per emotion category, the results from the cluster analyses slightly exceeded the expected 8-cluster-solution. However, requesting a single solution with 8 clusters grouped all vignettes according to their target emotion. To only include the most similar vignettes per emotion category, the vignette with the lowest hit rate per emotion category was excluded which led to one cluster per included emotion category for the Portuguese and English stimulus set in subsequent analyses. The German stimulus set included one second level cluster, because one disgust vignette did not reach as high disgust attributions as the other three disgust vignettes. However, the additional cluster occurred at the second level and between disgust vignettes themselves; the next cluster only occurred at the 22nd level. The single solution with specified 8 clusters again grouped all vignettes according to their target emotion. It can be concluded that the final stimulus set of 32 emotion vignettes includes the most distinct stimuli which map clearly onto specific emotion categories for all three languages. As it is general practice to include example stimuli in psychological research with the aim to familiarise participants with the task procedures, the 8 excluded emotion vignettes with the lowest hit rates per emotion category could be used for such purposes.

The individual dendrograms further showed that some emotion categories were more similar to each other than others. For example, the emotion categories of happiness and gratitude were positioned closer to each other than categories such as anger, guilt, and sadness, while anger was positioned a little farther from the other emotion categories. It seems as though emotion categories positive in valence and emotion categories negative in valence were each positioned closer together. In addition, emotions with higher arousal level were positioned closer to each other than such of low arousal. Such a structure is in line with emotion theories such as the circumplex model of affect (Russell, 1980) defining emotions as representable on valence

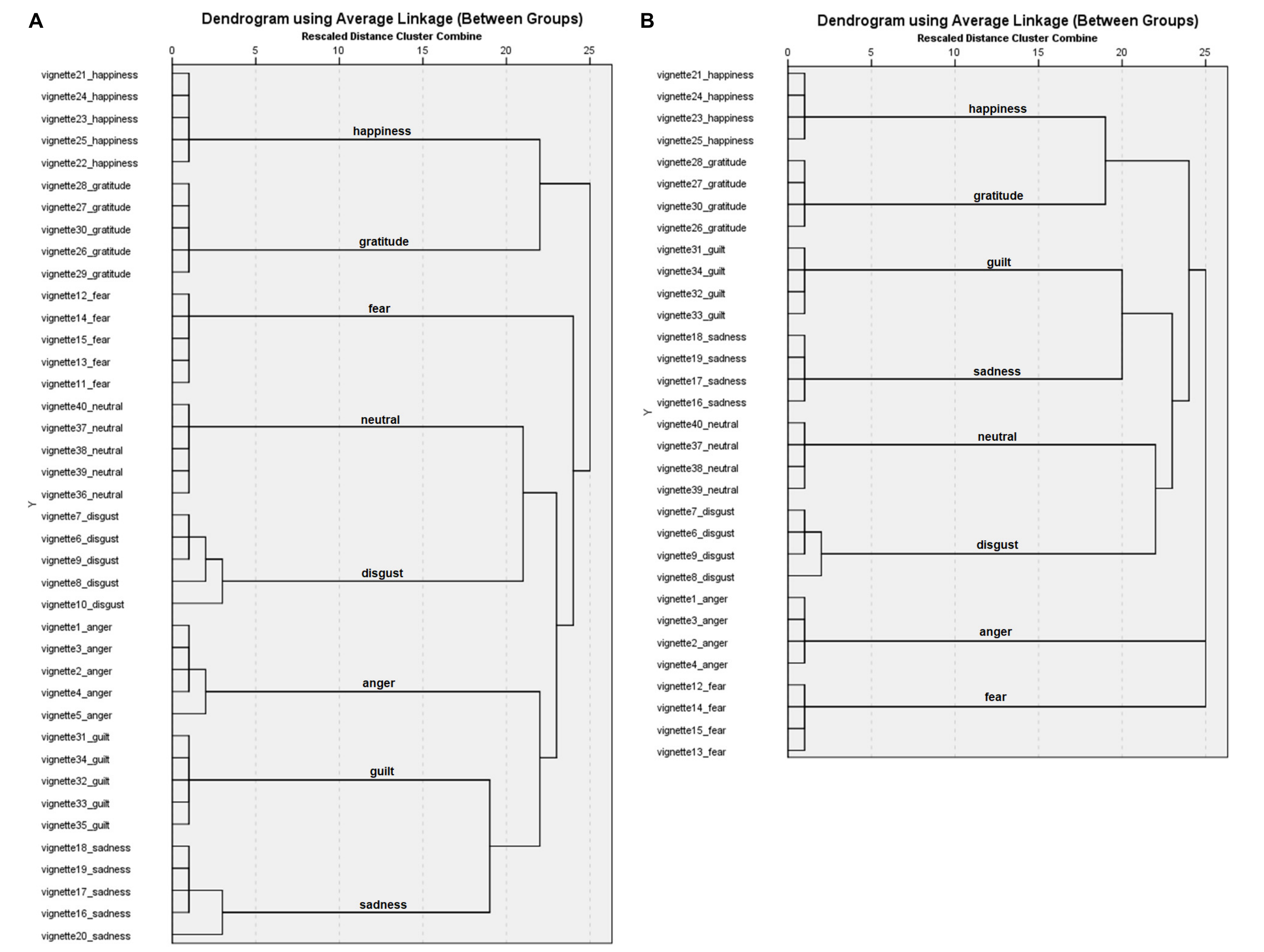


FIGURE 5 | Dendrograms for the German vignettes with **(A)** 40 vignettes and **(B)** 32 vignettes.

and arousal dimensions. When representing emotions in the two-dimensional space on valence and arousal, then negative emotion categories low in arousal are closer to each other (e.g., guilt and sadness) than to positive valence emotions that are low in arousal (e.g., happiness and gratitude), which themselves are closer to each other. It is interesting to note that the clustering in the current research was based on emotion label attributions of the emotion experienced while participants read scenarios rather than evaluations of the vignettes, e.g., on similarity. These results suggest that even when semantic understanding is necessary and a more cognitive approach to emotion elicitation is taken, the structure of emotion is represented. That is, it is more likely for participants to experience an emotion that is neighbouring the target emotion if it was not the target emotion that was experienced.

There were a few differences next to overlap between the three languages in terms of which individual emotion vignette per emotion category achieved the lowest hit rates (and was excluded from the main stimulus set per language). The neutral vignette with the lowest hit rate was different for all three languages. The lowest hit rate for anger and happiness vignettes were the same for the German and the English sample but not the Portuguese sample. However, the same vignettes led to lowest hit rates in all three languages for the emotion categories of fear, disgust, sadness, gratitude, and guilt. With many emotion categories overlapping in terms of which vignette had the lowest hit rate, this shows some consistency between the stimulus sets of the three languages.

The raw hit rates per emotion category were generally high and ranged between ∼75 and 95% in the Portuguese-speaking sample, ∼70–85% in the English-speaking sample, and ∼75–90% in the German-speaking sample. Even after correcting for response biases, the unbiased hit rates remained high in all three languages lowering the raw hit rates by roughly 5–10% per emotion category. Since there are no published verbal vignette stimulus sets including a similar number of emotion categories and the number of answer choices affects hit rates, the hit rates from the present stimulus sets cannot be directly compared to other stimulus sets. Nonetheless, these high agreement rates suggest that the stimulus sets in all three languages would

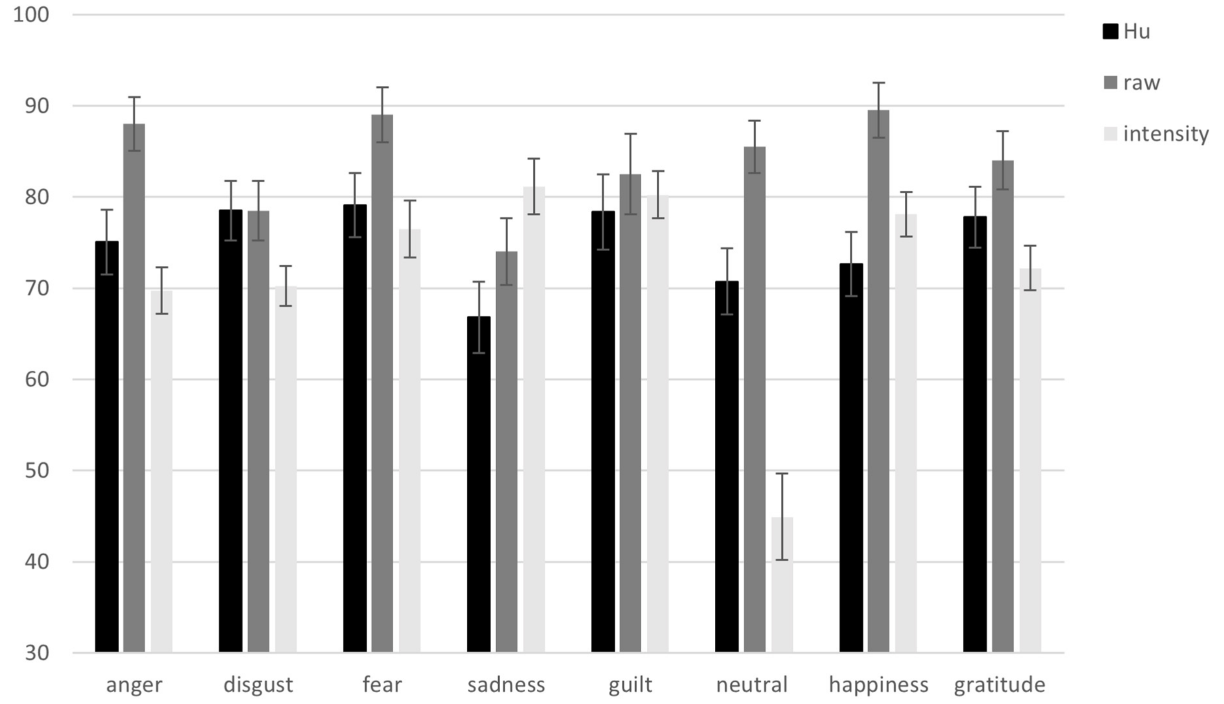


FIGURE 6 | Unbiased hit rates (*Hu*), raw hit rates, and intensity rates from the German vignettes validation per emotion category. Error bars represent standard errors of the means.

be suitable for application in emotion recognition research. High agreement on participants' reports about the emotion they experienced while immersing themselves into the scenarios described in the vignettes are also a prerequisite for applicability of the vignettes as valid emotion elicitation instrument.

The self-reported felt intensity reached medium to high intensities per emotion category suggesting that the vignettes are suitable for emotion elicitation. There were slight differences between the three languages regarding the intensity rates. The intensity rates (including the neutral category) in the Portuguese-speaking sample were ∼65–90%, ∼50–80% in the English-speaking sample, and ∼45–80% in the German-speaking sample. These results are only comparable to published film-based stimulus sets applicable for eliciting specific emotions, since no verbal vignette stimulus set is published presenting intensity ratings. Gross and Levenson (1995) reported between 37 and 64% intensity of felt emotions for the emotion categories included in their video stimulus set. The results from the vignettes presented here compare favourably to this stimulus set. The here obtained ranges of emotion intensity are below ceiling and thus allow for experimental manipulations aiming at investigating subsequent effects on emotion experience. For example, a study conducted in our laboratory showed that affiliative touch can modulate the evaluation of affective images (Wingenbach et al., unpublished). The created stimulus set could be used to investigate the effect of touch on emotion experience. Together, the created vignettes constitute a promising stimulus set for emotion elicitation.

There were differences in the hit rates between the three samples and the Portuguese sample achieved the highest hit rates across emotion categories. The samples differed from each other in their demographic characteristics, which can likely explain the differences in hit rates. The Portuguese sample included only university students who are required to participate in research as part of their degree and thus might have had prior experience with tasks as the current one. Better task performance by university students is often observed compared to general population samples and might also apply to the current research. In addition, the student sample included younger participants than the general population samples and the vignettes were written by age-similar peers. It is possible that these factors contributed to the higher hit rates in the Portuguese sample. Due to the differences between the samples, statistical comparisons of the results between the samples were not conducted.

In conclusion, three stimulus sets containing 32 vignettes (4 vignettes for each category of anger, disgust, fear, sadness, happiness, gratitude, guilt, and neutral) and an additional practice vignette per category were created and validated in three languages (Portuguese, English, and German) and the results suggest their suitability for emotion recognition and emotion elicitation research. The vignettes can be used for research purposes and are available to researchers free of charge downloadable from the **Supplementary Material**.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Mackenzie Presbyterian University Ethics

Committee' with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Mackenzie Presbyterian University Ethics Committee'.

AUTHOR CONTRIBUTIONS

PB conceptualised the study. LM and AH wrote the vignettes and collected the data. TW performed the data analysis and wrote the first version of the manuscript. All authors contributed to the data interpretation, manuscript writing, and approved the final version of the manuscript for submission.

FUNDING

This research was supported by the São Paulo Science Foundation (FAPESP) and Natura Cosméticos S.A. (Grant Nos. 2014/50282-5 and 2017/10501-8) including individual fellowships to TW (2017/00738-0), LM (2016/19277-0), and AH (2016/19167-0). PB was supported by both FAPESP and Natura Cosméticos S.A. (Grant Nos. 2014/50282-5 and 2017/10501-8) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (Grant No. 311641/2015-6).

ACKNOWLEDGMENTS

We thank Rosanna K. Smith for her help in recruiting native English-speaking participants and Fanny Lachat for her initial work in the project. We also thank everyone for participating.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01135/full#supplementary-material

TABLE S1 | All 40 vignettes for each of the 3 languages.

REFERENCES

Bänziger, T., Grandjean, D., and Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: the multimodal emotion recognition test (MERT). *Emotion* 9, 691–704. doi: 10.1037/a0017088

Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion* 12, 1161–1179. doi: 10.1037/a0025827

Belin, P., Fillion-Bilodeau, S., and Gosselin, F. (2008). The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behav. Res. Methods* 40, 531–539. doi: 10.3758/BRM.40.2.531

Carvalho, S., Leite, J., Galdo-Álvarez, S., and Gonçalves, ÓF. (2012). The emotional movie database (EMDB): a self-report and psychophysiological study. *Appl. Psychophysiol. Biofeedback* 37, 279–294. doi: 10.1007/s10484-012-9201-6

Dyck, M. J. (2012). The ability to understand the experience of other people: development and validation of the emotion recognition scales. *Aust. Psychol.* 47, 49–57. doi: 10.1111/j.1742-9544.2011.00047.x

Ekman, P., and Cordaro, D. (2011). What is meant by calling emotions basic. *Emot. Rev.* 3, 364–370. doi: 10.1177/1754073911410740

Ekman, P., and Friesen, W. (1976). *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Sorenson, E. R., and Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotion. *Science* 164, 86–88. doi: 10.1126/science.164.3875.86

Gross, J. J., and Levenson, R. W. (1995). Emotion elicitation using films. *Cogn. Emot.* 9, 87–108. doi: 10.1080/02699939508408966

Hareli, S., Sharabi, M., and Hess, U. (2011). Tell me who you are and I tell you how you feel: expected emotional reactions to success and failure are influenced by knowledge about a person's personality. *Int. J. Psychol.* 46, 310–320. doi: 10.1080/00207594.2010.547583

Hawk, S. T., van Kleef, G. A., Fischer, A. H., and van der Schalk, J. (2009). "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion* 9, 293–305. doi: 10.1037/a0015178

IBM Corp (2016). *Released. IBM SPSS Statistics for Windows, Version 24.0.* Armonk, NY: IBM Corp.

Ito, T. A., Cacioppo, J. T., and Lang, P. J. (1998). Eliciting affect using the international affective picture system: trajectories through evaluative space. *Personal. Soc. Psychol. Bull.* 24, 855–879. doi: 10.1177/0146167298248006

Krumhuber, E. G., Kappas, A., and Manstead, A. S. R. (2013). Effects of dynamic aspects of facial expressions: a review. *Emot. Rev.* 5, 41–46. doi: 10.1177/1754073912451349

Lane, J., and Anderson, N. H. (1976). Integration of intention and outcome in moral judgment. *Mem. Cogn.* 4, 1–5. doi: 10.3758/BF03213247

Lang, P. J., Bradley, M. M., and Cuthbert, B. N. (1997). *International Affective Picture System (IAPS): Technical Manual and Affective Ratings*. Gainesville, FL: NIMH Center for the Study of Emotion and Attention, 39–58.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., and van Knippenberg, A. (2010). Presentation and validation of the radboud faces database. *Cogn. Emot.* 24, 1377–1388. doi: 10.1080/02699930903485076

MacCann, C., and Roberts, R. D. (2008). New paradigms for assessing emotional intelligence: theory and data. *Emotion* 8, 540–551. doi: 10.1037/a0012746

McCullough, M. E., Kimeldorf, M. B., and Cohen, A. D. (2008). An adaptation for altruism. *Curr. Dir. Psychol. Sci.* 17, 281–285. doi: 10.1111/j.1467-8721.2008.00590.x

McHugo, G. J., Smith, C. A., and Lanzetta, J. T. (1982). The structure of self-reports of emotional responses to film segments. *Motiv. Emot.* 6, 365–385. doi: 10.1007/BF00998191

Ortony, A., and Turner, T. J. (1990). What's basic about basic emotions? *Psychol. Rev.* 97, 315–331. doi: 10.1037/0033-295X.97.3.315

Philippot, P. (1993). Inducing and assessing differentiated emotion-feeling states in the laboratory. *Cogn. Emot.* 7, 171–193. doi: 10.1080/02699939308409183

Russell, J. A. (1980). A circumplex model of affect. *J. Personal. Soc. Psychol.* 39:1161. doi: 10.1037/h0077714

Schaefer, A., Nils, F., Sanchez, X., and Philippot, P. (2010). Assessing the effectiveness of a large database of emotion-eliciting films: a new tool for emotion researchers. *Cogn. Emot.* 24, 1153–1172. doi: 10.1080/02699930903274322

Schlegel, K., and Scherer, K. R. (2017). The nomological network of emotion knowledge and emotion understanding in adults: evidence from two new performance-based tests. *Cogn. Emot.* 32, 1514–1530. doi: 10.1080/02699931.2017.1414687

Tracy, J. L., and Robins, R. W. (2006). Appraisal antecedents of shame and guilt: support for a theoretical model. *Pers. Soc. Psychol. Bull.* 32, 1339–1351. doi: 10.1177/0146167206290212

Vine, V., Bernstein, E. E., and Nolen-Hoeksema, S. (2018). Less is more? Effects of exhaustive vs. minimal emotion labelling on emotion regulation strategy planning. *Cogn. Emot.* doi: 10.1080/02699931.2018.1486286 [Epub ahead of print].

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *J. Nonverb. Behav.* 17, 3–28.

Wingenbach, T. S. H., Ashwin, C., and Brosnan, M. (2016). Validation of the amsterdam dynamic facial expression set - bath intensity variations (ADFES-BIV): a set of videos expressing low, intermediate, and high intensity emotions. *PLoS One* 11:e0147112. doi: 10.1371/journal.pone.0147112

Wood, A. M., Maltby, J., Stewart, N., Linley, P. A., and Joseph, S. (2008). A social-cognitive model of trait and state levels of gratitude. *Emotion* 8, 281–290. doi: 10.1037/1528-3542.8.2.281

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 18 June 2019
doi: 10.3389/feduc.2019.00047

Multilevel Generalized Mantel-Haenszel for Differential Item Functioning Detection

Brian F. French[1]*, **W. Holmes Finch**[2]* **and Jason C. Immekus**[3]

[1] Department of Kinesiology and Educational Psychology, Washington State University, Pullman, WA, United States, [2] Department of Educational Psychology, Ball State University, Muncie, IN, United States, [3] Department of Educational Leadership, Evaluation and Organizational Development, University of Louisville, Louisville, KY, United States

OPEN ACCESS

Edited by:
Elisa Pedroli,
Istituto Auxologico Italiano
(IRCCS), Italy

Reviewed by:
Yong Luo,
Educational Testing Service,
United States
Raman Grover,
British Columbia Ministry of
Education, Canada

***Correspondence:**
Brian F. French
frenchb@wsu.edu
W. Holmes Finch
whfinch@bsu.edu

Specialty section:
This article was submitted to
Assessment, Testing and Applied
Measurement,
a section of the journal
Frontiers in Education

Received: 05 March 2019
Accepted: 10 May 2019
Published: 18 June 2019

Citation:
French BF, Finch WH and
Immekus JC (2019) Multilevel
Generalized Mantel-Haenszel for
Differential Item Functioning Detection.
Front. Educ. 4:47.
doi: 10.3389/feduc.2019.00047

Research has demonstrated that when data are collected in a multilevel framework, standard single level differential item functioning (DIF) analyses can yield incorrect results, particularly inflated Type I error rates. Prior research in this area has focused almost exclusively on dichotomous items. Thus, the purpose of this simulation study was to examine the performance of the Generalized Mantel-Haenszel (GMH) procedure and a Multilevel GMH (MGMH) procedure for the detection of uniform differential item functioning (DIF) in the presence of multilevel data with polytomous items. Multilevel data were generated with manipulated factors (e.g., intraclass correction, subjects per cluster) to examine Type I error rates and statistical power to detect DIF. Results highlight the differences in DIF detection when the analytic strategy matches the data structure. Specifically, the GMH had an inflated Type I error rate across conditions, and thus an artificially high power rate. Alternatively, the MGMH had good power rates while maintaining control of the Type I error rate. Directions for future research are provided.

Keywords: multilevel, differential item functioning, invariance, validity, test and item development

INTRODUCTION

Measurement invariance (MI) is recognized as a critical component toward building a validity argument to support test score use and interpretation in the context of fairness. At the item-level, MI indicates that the statistical properties characterizing an item (e.g., difficulty) are equivalent across diverse examinee groups (e.g., language). As such, it represents a critical aspect of the validity of test data, particularly for ensuring the comparability of item and total scores to guide decisions (e.g., placement) across examine groups. Differential item functioning (DIF) is a direct threat to the MI of test items and occurs when item parameters differ across equal ability groups, resulting in the differential likelihood of a particular (e.g., correct) item response (Raju et al., 2002). DIF detection generally focus on the identification of uniform and nonuniform DIF, where uniform DIF refers to differential item difficulty across equal ability groups, and nonuniform DIF refers to inequality of the discrimination parameters across groups, after matching on ability. DIF studies are encouraged by the *Standards for Educational and Psychological Tests* (American Educational Research Association et al., 2014), and follow sound testing practices.

Considerable attention has been focused on the development and evaluation of DIF detection methods to identify potentially biased test items (Osterlind and Everson, 2009). The outcome of this work, for example, has provided a basis to judge the efficacy of these methods to detect DIF among dichotomously (Holland and Thayer, 1988; Narayanan and Swaminathan, 1996) and polytomously (French and Miller, 1996; Williams and Beretvas, 2006; Penfield, 2007) scored items. An extension of this work is testing their effectiveness to detect DIF under multilevel data structures (Luppescu, 2002; French and Finch, 2010, 2012, 2013; Jin et al., 2014). Hierarchical data structures, such as students nested in classrooms, are common in educational testing settings (O'Connell and McCoach, 2008). Consequently, the non-independence of observations in multilevel data can result in inflated Type I error rates (Raudenbush and Bryk, 2002), which can result in invalid inferences of DIF detection methods. Whereas adjusted DIF detection procedures (e.g., Mantel-Haenszel [MH], logistic regression [LR]) have been evaluated for dichotomously scored test items (French and Finch, 2012, 2013; Jin et al., 2014), the purpose of this study was to address the literature gap on the use of the generalized Mantel-Haenszel (GMH) procedure for DIF detection of polytomously scored test items in multilevel data.

DIF ASSESSMENT FOR POLYTOMOUS ITEM RESPONSE DATA USING THE GENERALIZED MANTEL-HAENSZEL STATISTIC

There exist a large number of DIF detection methods for diverse types of item data, several of which have been studied and compared (e.g., Narayanan and Swaminathan, 1996; Penfield, 2001; Kistjansson et al., 2004; Finch, 2005; Woods, 2011; Oliveri et al., 2012; Jin et al., 2014). In the context of polytomous item response data, which is the focus of this study, one of the most proven of these methods is the GMH statistic. Holland and Thayer (1988), and Narayanan and Swaminathan (1996), applied the MH to DIF detection with dichotomous items. Subsequently, it has been used for investigating the presence of DIF with polytomous items, and been shown to be a useful tool for that purpose (Penfield, 2001). The MH procedure is an extension of the chi-square test of association, allowing for comparison of item responses between the focal and reference groups conditioning across multiple levels of a matching subtest score. When testing the null hypothesis of no DIF, the MHχ^2 statistic is used (Holland and Thayer, 1988):

$$\frac{\{|\sum_{j=1}^{S}[A_j - E(A_j)]| - .5\}^2}{\sum_{j=1}^{S} Var(A_j)}, \tag{1}$$

where

$$Var(A_j) = \frac{n_{Rj}n_{Fj}m_{1j}m_{0j}}{T^2 j(T_j - 1)}, \tag{2}$$

In Equations (1) and (2), $A_j - E(A_j)$ is the difference between the observed number of correct responses for the reference group on the item being studied for DIF (A) and the expected correct number, $n_{Rj}$and n_{Fj} are the sample sizes for the reference and focal group, respectively, at score j of the matching subtest, m_{1j} *and* m_{0j}represent the number of correct and incorrect responses, respectively, at j matching subtest score, and T represents the total number of examinees at matching subtest score j. This statistic is distributed as a chi-square with one degree of freedom and tests the null hypothesis of no uniform DIF. This statistic can be readily extended to accommodate items with more than two categories (Penfield, 2001).

ADJUSTED MH TEST STATISTIC METHOD

French and Finch (2013) identified a promising set of adjustments for the MH statistic for DIF detection in the context of multilevel data. Their work was based on an earlier effort by Begg (1999) who demonstrated how the standard MH test statistic could be adjusted to account for multilevel data. The Begg MH (BMH) technique is based on the observation that the score statistic obtained from logistic regression is equivalent to the MH test statistic when the intraclass correlation (ICC) is equal to 0 (see Begg, 1999). Therefore, the variance associated with the logistic regression score statistic is proportional to the variance of the MH test statistic used for DIF detection. Notably, it is the variance and standard error of the MH test statistic that is underestimated in the presence of multilevel data. Given this relationship between the score statistic MH variances, BMH adjusts the MH test statistic by the ratio of the score statistic variance estimated using a logistic regression model accounting for the multilevel data structure with the generalized estimating equation (GEE) to the naïve score statistic variance that does not account for the multilevel nature of the data. The naïve and GEE-based logistic regression models both take the form:

$$\begin{aligned}
&\ln\left(\frac{P_{ki}}{1-P_{ki}}\right) = \beta_0 + \beta_1 X_i + \beta_2 Y_i \\
&\textit{where}, \\
&P_{ki} = \text{probability of a correct response to item k} \\
&\beta_0 = \text{intercept} \\
&X_i = \text{group membership for subject i} \\
&Y_i = \textit{matching} \text{ subtest score for subject i} \\
&\beta_1 = \text{coefficient for group variable} \\
&\beta_2 = \text{coefficient for matching subtest variable}
\end{aligned} \tag{3}$$

For the naïve LR model, the covariance matrix for the dependent variable with respect to clusters is the identity matrix, in which the off-diagonal elements are 0, reflecting no clustering effects on the outcome (i.e., ICC = 0). The GEE model estimates the off-diagonal elements of the covariance matrix, thus accounting for within cluster correlations among responses. In this case, the unstructured covariance matrix is estimated, meaning that a unique covariance was estimated for each cluster. For both naïve LR and GEE, the variances of the score statistic are obtained and used to calculate their adjustment factor, which appears in

Equation (4) below.

$$
\begin{aligned}
&f = \tfrac{\sigma^2_{GEE}}{\sigma^2_{Naive}} \\
&\text{where,} \\
&\sigma^2_{GEE} = \text{GEE adjusted variance of the score statistic accounting for clustering} \\
&\sigma^2_{Naive} = \text{Naive variance of the score statistic ignoring clustering; proportional to the variance of MH}
\end{aligned}
\tag{4}
$$

If the ICC is 0 in the population, then this ratio will be near 1 for the sample. However, as the within cluster correlation among observations increases so does σ^2_{GEE}, f will also increase in value, reflecting the overestimation of the score statistic variance in the presence of multilevel data. The f ratio can then be used to adjust the MH test statistic as seen in Equation (5).

$$MH_B = \frac{MH}{f} \tag{5}$$

MH is the standard MH chi-square test statistic. As noted above, when the within-cluster correlations are large, σ^2_{GEE} will be larger than σ^2_{naive}, leading to a value of f that is relatively large and positive, which, will lead to a larger value of f, which when applied in Equation (5) will decrease the size of MH_B relative to MH. This will correct for the within cluster correlation induced by the multilevel data structure.

The use of the MH_B statistic for dichotomous DIF detection demonstrated that while it was very effective at controlling the Type I error rate in the presence of multilevel data, it exhibited markedly lower power for relatively small sample sizes, and lower levels of DIF (French and Finch, 2013). Thus, it was suggested that alternative adjustments to f be considered. These alternatives included multiplying f by 0.85 (BMH85), 0.90 (BMH9), or 0.95 (BMH95) to reduce the amount of the correction. These adjustments were selected through an iterative process of experimentation with the method, and validation using Monte Carlo simulations (French and Finch, 2013). Empirical results of the simulation study involving dichotomous data showed that the standard BMH statistic, as well as the BMH95 and BMH9 statistics, were able to maintain the nominal Type I error rate across all study conditions. However, they also demonstrated lower power than MH across many of these same data conditions. On the other hand, MH consistently displayed inflated Type I error rates in the presence of multilevel data for testing DIF with a between clusters variable. The BMH85 statistic offered a reasonable compromise for DIF in the presence of multilevel data, particularly when the ICC was 0.25 or greater given Type I error inflation never exceeded 0.093 (compared to Type I error rates in excess of 0.20 for MH), and it maintained power rates close to MH.

GOALS OF THE CURRENT STUDY

The goal of this study was to examine the performance of the Begg adjusted methods for MH in the context of polytomous item data and build upon the foundation laid with dichotomous items. Given that the GMH approach has been shown to be an effective DIF detection tool for polytomous data, it was of interest to ascertain how well an adjusted version of the statistic would work in the context of multilevel data, using the Begg adjustment based methods outlined above (i.e., BGMH85, BGMH9, and BGMH95). It was expected that BGMH85 would perform best of the options compared. Thus, the current simulation study examined the Type I error and power rates for DIF detection with polytomous items using GMH, BGMH85, BGMH9, and BGMH95 across manipulated factors (e.g., grouping variable, ICC, subjects per cluster).

METHODS

A simulation study (1,000 replications) using SAS (V9.3) compared the performance of the BGMH adjustments to standard GMH for DIF detection with polytomously scored items. Outcome variables of interest included Type I error and power rates across manipulated factors, including: grouping variable, ICC, number of clusters, sample size per cluster, and DIF magnitude. We note that the standard equation for the ICC is different for ordinal variables where the within variance is a constant (i.e., 3.29, Heck et al., 2013). Data were simulated using a multilevel graded response model (MGRM; e.g., Fox, 2005; Kamata and Vaughn, 2011), with item threshold parameters and discrimination values appearing in **Table 1**. The model can be defined using Kamata and Vaughn's general example:

$$P_{x_i}\left(\theta_{jk}, \theta_{.k}\right) = \frac{e^{\left(\alpha_i^{(s)}\theta_{jk} + \alpha_i^{(c)}\theta_{.k} - \delta_{x_i}\right)}}{1 + e^{\left(\alpha_i^{(s)}\theta_{jk} + \alpha_i^{(c)}\theta_{.k} - \delta_{x_i}\right)}} \tag{6}$$

Where

θ_{jk} = Latent trait for student j in cluster k or the amount of deviation from the group mean ability for student j in cluster k.
$\theta_{.k}$ = Latent trait for cluster k or group mean ability
$\alpha_i^{(s)}$ = Discrimination parameter for item i at student level
$\alpha_i^{(c)}$ = Discrimination parameter for item i at cluster level
δ_{x_i} = Threshold for item i for category boundary x

The latent traits are assumed to be distributed as follows:

$$\theta_{jk} \sim N\left(0, \sigma^2_{\theta^{(s)}}\right)$$
$$\theta_{.k} \sim N\left(0, \sigma^2_{\theta^{(c)}}\right)$$

This would give the probability of obtaining a certain score or higher and the probability of obtaining a certain category would be computed as the difference between this probability of x or higher and the probability of responding in category x + 1 or higher (e.g., Natesan et al., 2010; Kamata and Vaughn, 2011).

For all simulations, 20 items were simulated, each with 4 response levels, and a purified scale score was used for matching purposes. This latter condition was used to allow for the isolation of the impact of multilevel data, exclusive of other factors that might influence the performance of GMH and the adjustments

TABLE 1 | Data generating parameters for the graded response model.

Item	Discrimination	T1	T2	T3
1	0.89	−1.22	0	1.37
2	1.03	−1.50	−0.67	1.19
3	0.78	−1.41	0.14	1.20
4	1.44	−0.87	0.5	1.06
5	1.71	−1.87	0.89	1.49
6	0.99	−1.16	−0.29	1.13
7	1.36	−0.89	0.35	0.87
8	1.05	−1.09	0.2	1.58
9	1.29	−1.14	0.22	1.64
10	1.65	−1.25	0.17	1.46
11	0.88	−1.00	0.32	1.38
12	0.93	−1.75	−0.59	1.34
13	1.04	−0.77	0.08	1.49
14	0.91	−1.81	0.22	1.15
15	1.55	−1.10	0.04	1.98
16	0.87	−1.16	−0.29	1.13
17	1.32	−0.89	0.35	0.87
18	1.47	−1.09	0.20	1.58
19	0.90	−1.14	0.22	1.64
20	1.63	−1.25	0.17	1.46

(e.g., contaminated scale). DIF was simulated for a target item, with magnitudes as described below. In the calculation of the MH statistics, purified raw test scores were used for matching purposes.

MANIPULATED FACTORS

Grouping Variable

Two grouping variable conditions were simulated: (1) within-cluster (e.g., examinee gender), or (2) between-cluster (e.g., teaching method, teacher gender), consistent with previous research on DIF detection within multilevel data structures (French and Finch, 2013; Jin et al., 2014).

Intraclass Correlation (ICC)

For the studied item and total score, the ICCs were set at five levels: 0.05, 0.15, 0.25, 0.35, and 0.45. These values were in accord with estimates obtained from large national databases (Hedges and Hedberg, 2007), and reflect values observed in practice (Muthén, 1994).

Number of Clusters

The number of simulated level-2 clusters included: 50, 100, and 200. Prior studies (Muthén and Satorra, 1995; Hox and Maas, 2001; Maas and Hox, 2005; French and Finch, 2013) have used similar values.

Number of Subjects Per Cluster

Clusters were simulated to be of equal size, taking the values 5, 15, 25, and 50. These values match those used in previous research (Muthén and Satorra, 1995; Hox and Maas, 2001; Maas and Hox, 2005; French and Finch, 2013).

DIF Magnitude

Four levels of DIF magnitude were simulated for the target item, based on prior DIF simulation for polytomous items (Penfield, 2007), and included: 0, 0.4, 0.6, and 0.8. Uniform DIF was specified by simulating differences in item each threshold parameter value for the target item, between the groups. In other words, the DIF magnitude value was added to each of the threshold values (**Table 1**) on the target item for the focal group. The focus was on uniform DIF as the MH procedure is not accurate with non-uniform DIF. In addition, uniform DIF tends to occur with greater frequency in assessments compared to non-uniform DIF, as reflected in simulation work (Jodoin and Gierl, 2001; French and Maller, 2007), and applied work (e.g., Maller, 2001). Each replicated dataset per condition was analyzed using standard GMH and the MGMH methods outlined above.

Analysis

To determine which manipulated factors influenced the power and Type I error rates, repeated measures analysis of variance (ANOVA) was used, per recommendations for simulation research (Paxton et al., 2001; Feinberg and Rubright, 2016). A separate such analysis was conducted in which the Type I error or power rates averaged across replications for each combination of conditions served as the dependent variables. The manipulated factors described above, and their interactions, served as the independent variables in the model. In addition to statistical significance of these model terms, the η^2 effect size was also reported. We also focus on a visual display of the results to enhance comprehension and efficiency (McCrudden et al., 2015) compared to displaying many tables.

RESULTS

Type I Error Rate

The ANOVA results identified two terms significantly related to the Type I error rate of the GMH and Begg adjusted procedures. These included the 3-way interaction of the test statistic by ICC by grouping variable for which DIF was tested [$F_{(12,219)} = 33.749$, $p < 0.001$, $\eta^2 = 0.646$], and the 3-way interaction of test statistic by cluster size by grouping variable for which DIF was tested [$F_{(12,219)} = 8.752$, $p < 0.001$, $\eta^2 = 0.324$). **Figure 1** shows the Type I error rates of the statistical tests by the ICC and the grouping variable being tested for DIF. When this variable was at the within-cluster level (e.g., gender), the Type I error rate of the GMH test adhered to the nominal 0.05 level, regardless of the size of the ICC. Similarly, error rates of the Begg adjusted statistics were conservative, fell below the 0.05 level, and were not affected by ICC level. For the between-cluster grouping variable, GMH had inflated Type I error rates well beyond the 0.05 level and increased with ICC values. For the Begg adjusted values, Type I error rates increased slightly across ICC conditions but, nonetheless, were at or below the nominal level.

Figure 2 displays the Type I error rates for each statistical test by cluster size and grouping variable. As shown, when the grouping variable was within-cluster, the Type I error rates of all statistical methods, including the standard GMH, were at or below the nominal level of 0.05. For the Begg corrected

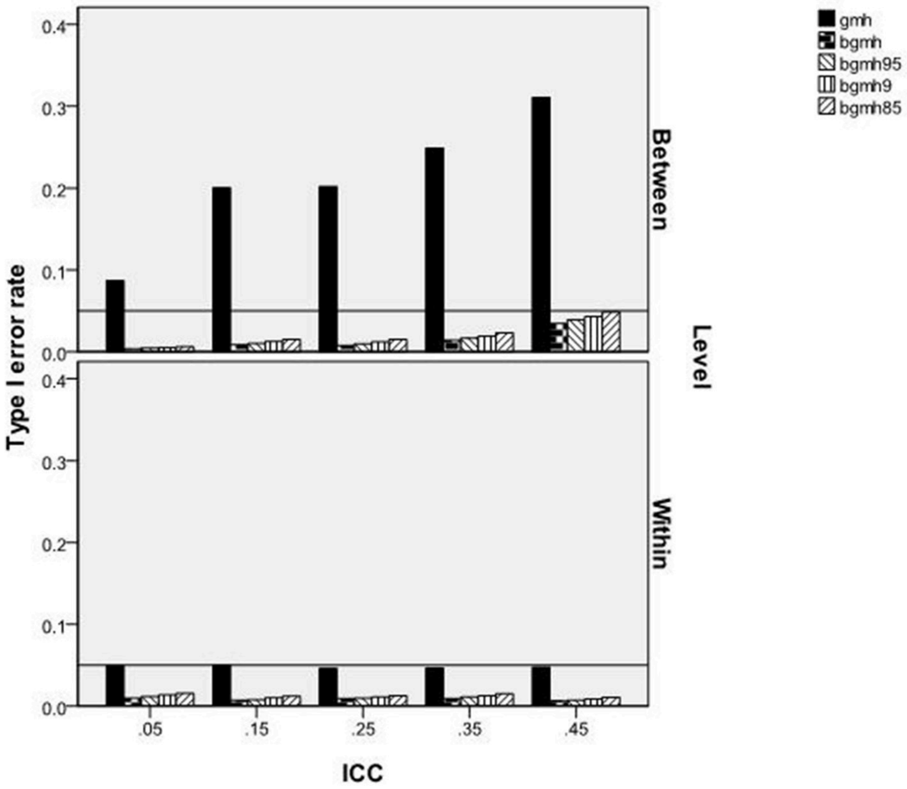


FIGURE 1 | Type I error rates of GMH and BGM test statistics by ICC and level of variable.

tests, the error rate was always below 0.05, and declined with increases in the sample size per cluster. In contrast, when the variable was between-cluster, the Type I error rate for GMH was always greater than the 0.05 level, and increased concomitantly with increases in sample size per cluster. Contrary, the Begg corrected tests maintained error rates below the 0.05 level and decreased with increases in the sample size per cluster.

Power

As with the Type I error rate, a repeated measures ANOVA was used to identify the significant main effects and interactions of the manipulated factors in terms of their impact on power rates. The interaction of ICC by method [$F_{(16,1,160)} = 6.147, p < 0.001, \eta^2 = 0.078$], the interaction of level of variable by amount of DIF by method [$F_{(8,576)} = 15.368$, $p < 0.001$, $\eta^2 = 0.176$], and the interaction of number of clusters by sample size per cluster by method [$F_{(24,1,160)} = 4.492$, $p < 0.001$, $\eta^2 = 0.085$] were each significantly related to power.

Table 2 reports power rates by method and ICC. Importantly, given the inflated Type I error rates in the between-cluster variable condition, power results for GMH must be interpreted with caution. Only when the ICC = 0.05 were the power rates for the Begg adjusted methods >0.80. Consequently, across the test statistics, power to detect DIF decreased with higher ICC values. Specifically, for the standard GMH, the decline in power from an ICC of 0.05 to 0.45 was approximately 0.045, whereas the Begg adjusted methods decline was 0.11.

Figure 3 reports power rates by the level of the variable (between, within), amount of DIF, and statistical test. As shown, for each test statistic, power increased concomitantly with increases in the amount of DIF present in the data. Furthermore, power rates were lower for the between-levels variable for all methods, except for GMH with DIF = 0.80, in which case power was approximately 1.0 across conditions. The GMH statistic had a distinct power advantage over the Begg adjusted methods for between- and within-level variables when DIF = 0.40, and for between-level variables when DIF = 0.60. At the two highest DIF levels, power for BGMH85 (the adjusted method with the highest power rates) was approximately equal to that of GMH for the within-cluster variable. However, power for all of the adjusted methods was at least 0.07 lower than that of GMH in the between-cluster variable condition. As previously noted, however, power rates for GMH in the between-cluster condition must be interpreted with caution, due to inflated Type I error rates.

Figure 4 displays power rates by statistical test, number of clusters, and sample size by cluster. Again, given the Type I error inflation for GMH that was reported earlier, these results must be interpreted with caution. For all of the methods studied here, power was higher with larger sample sizes and, for most conditions, power was greater for GMH when compared to

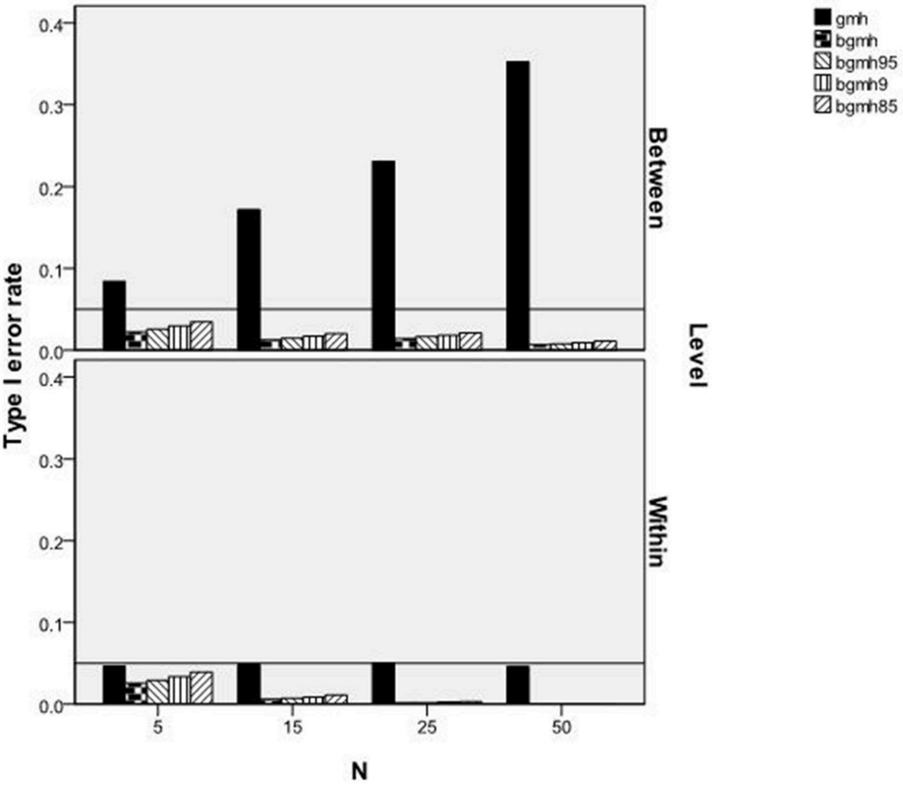


FIGURE 2 | Type I error rates of GMH and BGM test statistics by sample size per cluster and level of variable.

TABLE 2 | Power by method and ICC.

ICC	GMH	BGMH	BGMH95	BGMH9	BGMH85
0.05	0.907	0.788	0.800	0.811	0.823
0.15	0.895	0.764	0.776	0.787	0.799
0.25	0.895	0.752	0.762	0.776	0.789
0.35	0.880	0.723	0.728	0.744	0.757
0.45	0.862	0.673	0.682	0.700	0.710

the Begg adjusted methods. In addition, with more clusters the difference in power between GMH and the adjusted methods declined. For example, for the 100 clusters with 25 members per cluster condition and the 50 clusters with 50 members per cluster, both had a total sample size of 2,500. In both conditions, power for the GMH statistics was ∼0.98. However, for the Begg adjusted methods, the power in the 50 clusters condition was ∼0.20 lower than in the 100 clusters condition, despite that the total sample sizes for the two cases were identical. Indeed, for the 100 clusters with 25 members per cluster case, the power for BGMH85 was 0.08 lower than that of GMH, whereas it was 0.27 lower in the 50 clusters with 50 members per cluster condition. This example demonstrates the nature of the interaction among method, number of clusters, and cluster size; namely, that with more clusters the power of the Begg adjusted methods was greater, regardless of total sample size. Finally, in the presence of 200 clusters, the difference in power rates of the GMH and Begg adjusted methods were always <0.05, regardless of cluster size.

DISCUSSION

The goal of this study was to investigate the performance of the GMH and adjusted Begg methods for the detection of uniform DIF for polytomous test items in the presence of multilevel data. As such, it sought to extend the availability of DIF procedures to the context of multilevel data gathered on examinees grouped in clusters (e.g., classrooms, schools). The availability of multilevel statistical procedures ensures that analyses align with the data structure to ensure valid inferences to guide decisions (Raudenbush and Bryk, 2002; O'Connell and McCoach, 2008). Screening educational tests for DIF is an important step toward ensuring the accuracy of inferences based on between-group score differences within (e.g., language) and/or between clusters (e.g., schools). Furthermore, it is a critical step toward promoting fair testing practices in that tests function similarly across diverse examinee groups (American Educational Research Association et al., 2014). Therefore, it is crucial that appropriate DIF detection procedures exist to identify items that perform differentially for subgroups, when item response data are collected in a multilevel framework.

The Type I error rates of GMH and the Begg adjusted methods differed according to the manipulated factors. In particular, the

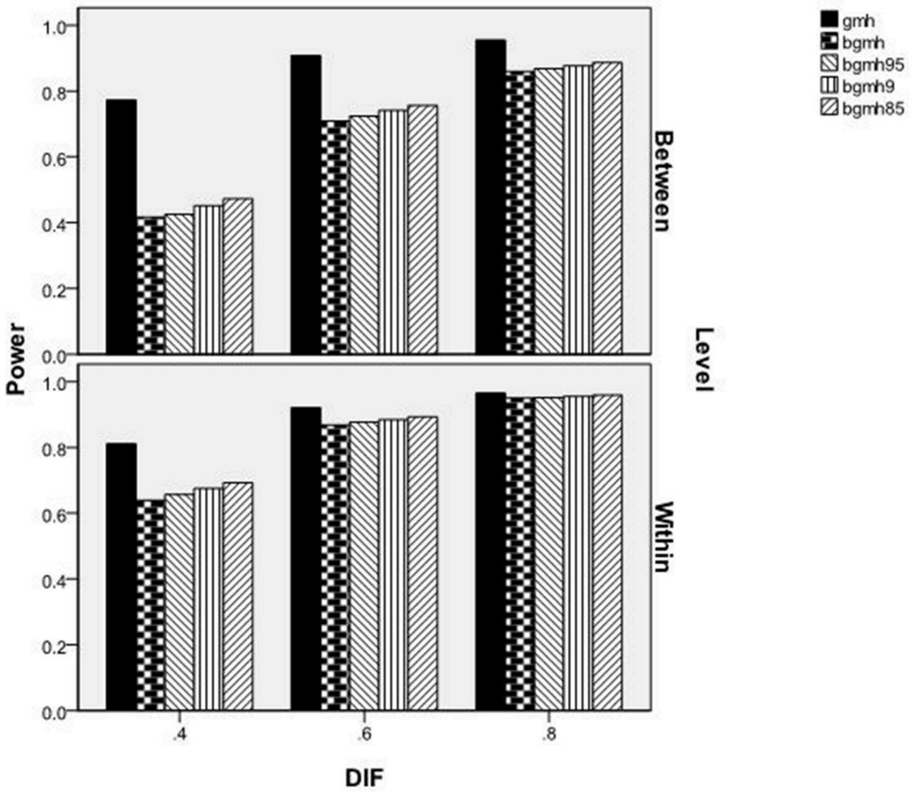


FIGURE 3 | Power by method, amount of DIF, and level of variable.

statistical significance of separate 3-way interactions indicated that the GMH procedure had inflated Type I error rates for specific conditions, whereas the Begg adjusted methods were more conservative and, in general, adhered to the nominal alpha level. Specifically, the procedures differed based on the grouping variable and ICC. For the within-cluster condition, all procedures reported Type I error rates at or below the nominal level, with the Begg adjusted methods being slightly more conservative than the GMH procedure. When the grouping variable was between-cluster (e.g., examinee gender), the collection of Begg adjusted methods reported acceptable Type I errors rates, whereas the GMH method was considerably more liberal. Notably, the Type I error rates for all procedures increased with associated increases of the ICC. The methods were also found to differ when combined with the grouping variable and number of subjects per cluster. As previously reported, when the grouping variable was within-cluster, all procedures adhered to the nominal 0.05 error rate, although the GMH procedure was slightly higher than the Begg adjusted methods. Additionally, the Type I error rates were found to decrease as the number of subjects per cluster increased. Conversely, when the grouping variable was between-cluster (e.g., schools assigned to different treatment conditions), the GMH procedure reported inflated Type I errors and increased when the number of subjects per cluster increased. On the other hand, the Begg adjusted methods adhered to the nominal 0.05 level, with their Type I error rates decreasing as the number of subjects per cluster increased. These findings contribute to the body of literature that standard DIF procedures (MH, LR) have inflated Type I error rates in the presence of multilevel data (Jin et al., 2014).

The statistical power of the GMH and Begg adjusted methods were also found to vary depending on manipulated factor. Although the statistical power of the GMH procedure exceeded 0.80 across ICC levels, it should be interpreted with great caution due to its inflated Type I error rates. Therefore, in the presence of multilevel data, the GMH procedure would be expected to erroneously report the presence of DIF among test items. Only when the ICC was 0.05 did the Begg adjusted methods report power estimates above the desired 0.80 level. As the variance associated to the cluster increases (ICCs 0.05–0.45), the statistical power of the methods decreased approximately 0.11 across the Begg adjusted procedures. Power rates also varied by level of the grouping variable (within or between) and amount of DIF. Notably, regardless of level of variable, power rates were lowest for the lowest level of DIF condition (i.e., 0.40), whereas GMH power was near 0.80. Again, despite the GMH procedure yielding power at or above 0.80 across conditions, the corresponding Type I error rates demand cautious interpretation. For both the within- and between-cluster conditions, power rates of the Beggs adjusted methods increased approximately to or above 0.80. Only

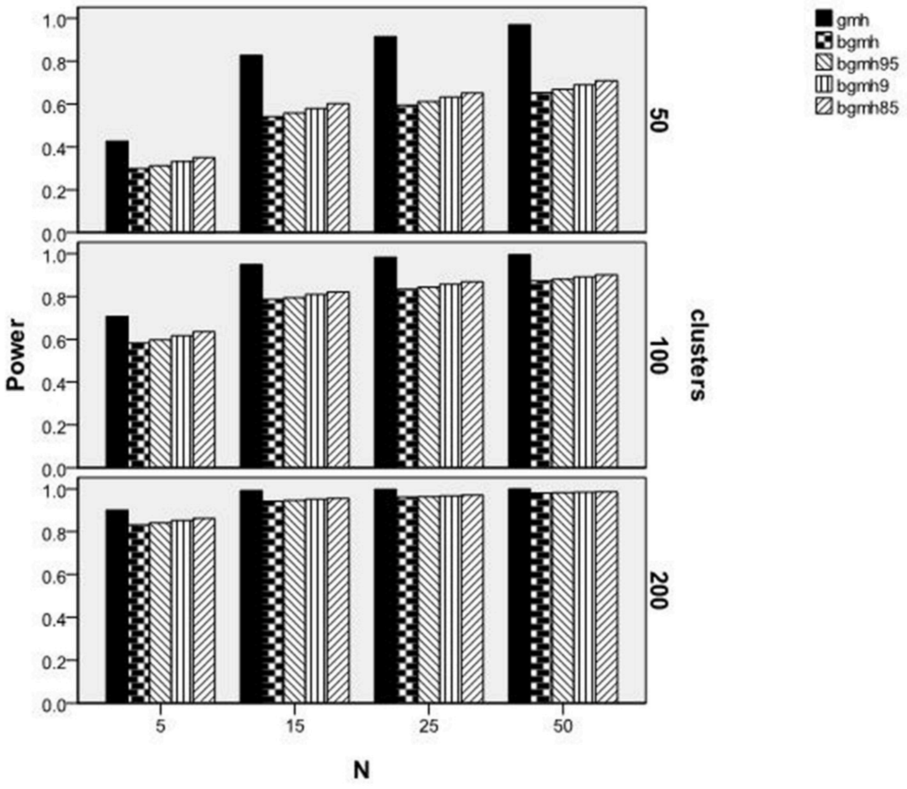


FIGURE 4 | Power by method, number of clusters, and sample size per cluster.

when the DIF magnitude was 0.60 did the Begg methods report statistical power above 0.80, irrespective of the grouping variable. Finally, across GMH procedures, power rates increased with the number of clusters (e.g., 50, 100) and the number of subjects per cluster. Notably, all procedures reported power rates <0.50 with 50 clusters and five subjects per cluster. Only when the number of clusters was 100 or 200 did the Begg methods report an acceptable level of power for DIF detection.

Empirical findings of the current study provide a framework for the application of the GMH and Begg adjusted procedures for DIF detection. In applied settings, the GMH procedure should be restricted for consideration in the absence of multilevel data. Even with an ICC of 0.05 and at the between-level, its Type I error rate was ∼0.10. This is similar to results with the MH and logistic regression procedures which are less precise in identifying DIF in multilevel data structures (French and Finch, 2010, 2013; Jin et al., 2014), particularly at the between-group level. On the other hand, the Begg adjusted values have generally reasonable power (>0.67) to detect DIF under varying multilevel conditions while maintaining an error rate at the nominal 0.05 level. One caveat is that when the number of clusters may be small (50 or less) and the sample size per cluster is also small, power for the Begg methods was found to be attenuated. Therefore, the collective set of Begg adjusted methods examined in this study seem most favorable for multilevel level data, although their power rates are expected to be slightly lower when the number of clusters is smaller. Study findings also provide a basis for ongoing investigations of DIF procedures under various conditions that may be found in applied testing contexts. For example, Jin et al. (2014) extended the work of French and Finch (2010, 2013) regarding the performance of hierarchical LR, LR, and MH under multilevel data structures when the ICC of the item was less than the ICC of the latent trait, in addition to other manipulated factors (e.g., item type, model type).

The confluence of results supports the need for continued research to identify DIF procedures that are accurate at identifying various types of DIF items under various multilevel structures expected in applied testing settings. For the practitioner, this work should allow one to screen for DIF items when multilevel data are present while maintaining control of Type I error and having adequate power to detect DIF. This increase in DIF accuracy, due to analyses matching the data structure, should guard against resources being wasted on reviewing items for problems as a result of in inflated error rate if an adjustment was not employed. In addition, software to implement these methods easily is needed. A SAS package with an easy to use interface is available from the authors for the Begg method for the dichotomous conditions. SAS and R packages are in development, which move the ideas presented here through simulation into practice.

This study contributes to the literature on the effectiveness of adjusted statistical methods for DIF detection in the presence of multilevel data. In particular, under multilevel data structures, the Begg adjusted methods performed most favorably in the detection of DIF for polytomous items. Nonetheless, the extent to which the methods examined in this study compare to other DIF detection methods proposed for polytomously scored items (e.g., French and Miller, 1996; Penfield, 2008) within a multilevel framework offers directions for continued research. Likewise, the manipulated factors examined represent a step toward examining additional factors that may contribute to the functioning of these methods in applied settings. The development and evaluation of DIF detection methods with multilevel data will contribute to the psychometric tools available to ensuring accurate item and total test scores to guide test-based decisions.

DATA AVAILABILITY

The datasets for this manuscript are not publicly available because these were simulated datasets. They can be reproduced. Requests to access the datasets should be directed to frenchb@wsu.edu.

AUTHOR CONTRIBUTIONS

BF was responsible for conceptualization of the idea, design, and conducting the study. WF was responsible for conceptualization of the idea, design, and conducting the study. JI was responsible for assisting with the review of the literature, editing, and quality control.

FUNDING

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D110014 to Washington State University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Begg, M. D. (1999). Analyzing k(2 × 2) tables under cluster sampling. *Biometric* 55, 302–307.

Feinberg, R. A., and Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educ. Meas. Issues Pract.* 36–49. doi: 10.1111/emip.12111

Finch, H. (2005). The MIMIC method as a method for detecting DIF: comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Appl. Psychol. Meas.* 29, 278–295. doi: 10.1177/0146621605275728

Fox, J. P. (2005). "Multilevel IRT model assessment," in *New Developments in Categorical Data Analysis for the Social and Behavioral Sciences*, eds L. A. van der Ark, M. A. Croon, and K. Sijtsma (New York, NY: Taylor & Francis), 227–252.

French, A. W., and Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning polytomous items. *J. Educ. Meas.* 33, 315–332. doi: 10.1111/j.1745-3984.1996.tb00495.x

French, B.F., and Finch, W. H. (2010). Hierarchical logistic regression: accounting for multilevel data in DIF detection. *J. Educ. Meas.* 47, 299–317. doi: 10.1111/j.1745-3984.2010.00115.x

French, B. F., and Finch, W. H. (2012). *April.* "Extensions of Mantel-Haenszel for multilevel DIF detection," in *Paper Presented at the American Educational Research Association Conference* (Vancouver, BC).

French, B. F., and Finch, W. H. (2013). Extensions of the Mantel-Haenszel for multilevel DIF detection. *Educ. Psychol. Meas.* 73, 648–671. doi: 10.1177/0013164412472341

French, B. F., and Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educ. Psychol. Meas.* 67, 373–393. doi: 10.1177/0013164406294781

Heck, R. H., Thomas, S., and Tabata, L. (2013). *Multilevel Modeling of Categorical Outcomes Using IBM SPSS.* New York, NY: Routledge. doi: 10.4324/9780203808986

Hedges, L. V., and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educ. Eval. Policy Anal.* 29, 60–87. doi: 10.3102/0162373707299706

Holland, P. W., and Thayer, D. T. (1988). "Differential item performance and the Mental-Haenszel procedure," in *Test Validity*, eds H. Wainer and H. I. Braun (Hillsdale, NJ: Lawrence Erlbaum), 129–145.

Hox, J. J., and Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Struct. Eq. Model.* 8, 157–174. doi: 10.1207/S15328007SEM0802_1

Jin, Y., Myers, N. D., and Ahn, S. (2014). Complex versus simple modeling for DIF detection: When the intraclass correlation coefficient (ρ) of the studied item is less than the ρ of the total score. *Educ. Psychol. Meas.* 74, 163–190. doi: 10.1177/0013164413497572

Jodoin, M. G., and Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Appl. Meas. Educ.* 14, 329–349. doi: 10.1207/S15324818AME1404_2

Kamata, A., and Vaughn, B. K. (2011). "Multilevel item response theory modeling," in *Handbook of Advanced Multilevel Analysis*, eds J. Hox and J. K. Roberts (New York, NY: Routledge), 41–57.

Kistjansson, E., Aylesworth, R., McDowell, I., and Zumbo, B. (2004). A comparison of four methods for detecting differential item functioning in ordered response items. *Educ. Psychol. Meas.* 65, 935–953. doi: 10.1177/0013164405275668

Luppescu, S. (2002). "DIF detection in HLM," in *Paper Presented at the Annual Meeting of the American Educational Research Association* (New Orleans, LA).

Maas, C. J. M., and Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology* 1, 86–92. doi: 10.1027/1614-2241.1.3.86

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educ. Psychol. Meas.* 61, 793–817. doi: 10.1177/00131640121971527

McCrudden, M. T., Schraw, G., and Buckendahl, C. (eds.) (2015). *Use of Visual Displays in Research and Testing: Coding, Interpreting, And Reporting Data* (Charlotte, NC: Information Age Publishing).

Muthén, B.O. (1994). Multilevel covariance structure analysis. *Sociol. Methods Res.* 22, 376–398. doi: 10.1177/0049124194022003006

Muthén, B.O., and Satorra, A. (1995). Complex survey data in structural equation modeling. *Sociol. Methodol.* 25, 267–316. doi: 10.2307/271070

Narayanan, P., and Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Appl. Psychol. Meas.* 20, 257–274. doi: 10.1177/014662169602000306

Natesan, P., Limbers, C., and Varni, J. W. (2010). Bayesian estimation of graded response multilevel models using Gibbs sampling: formulation and illustration. *Educ. Psychol. Meas.* 70, 420–439. doi: 10.1177/0013164409355696

O'Connell, A. A., and McCoach, D. B. (eds.) (2008). *Multilevel Modeling of Educational Data* (Charlotte, NC: Information Age Publishing).

Oliveri, M. A., Olson, B. F., Ercikan, K., and Zumbo, Z. (2012). Methodologies for investigating item and test-level measurement equivalence in international large-scale assessments. *Int. J. Testing* 12, 203–223. doi: 10.1080/15305058.2011.617475

Osterlind, S. J., and Everson, H. T. (2009). *Differential Item Functioning*, 2nd Edn. Thousand Oaks, CA: Sage.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., and Chen, F. (2001). Monte Carlo experiments: design and implementation. *Struct. Equat. Model.* 8, 287–312. doi: 10.1207/S15328007SEM0802_7

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: a comparison of three Mantel-Haenszel procedures. *Appl. Meas. Educ.* 14, 235–259. doi: 10.1207/S15324818AME1403_3

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *J. Educ. Meas.* 44, 187–210. doi: 10.1111/j.1745-3984.2007.00034.x

Penfield, R. D. (2008). Three classes of nonparametric differential step functioning effect estimators. *Appl. Psychol. Meas.* 32, 480–501.

Raju, N. S., Laffitte, L. J., and Byrne, B. M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *J. Appl. Psychol.* 87, 517–529.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd edn*. Thousand Oaks, CA: Sage.

Williams, N. J., and Beretvas, N. S. (2006). DIF identification using HGLM for polytomous items. *Appl. Psychol. Meas.* 30, 22–42. doi: 10.1177/0146621605279867

Woods, C. M. (2011). DIF testing for ordinal items with poly-SIBTEST, the Mantel and GMH tests, and IRT-LR-DIF when the latent distribution is nonnormal for both groups. *Appl. Psychol. Meas.* 35, 145–164. doi: 10.1177/0146621610377450

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 June 2019
doi: 10.3389/fpsyg.2019.01360

Factor Structure and Measurement Invariance Across Gender Groups of the 15-Item Geriatric Depression Scale Among Chinese Elders

*Haofei Zhao, Jiayue He, Jinyao Yi and Shuqiao Yao**

Medical Psychological Center, Second Xiangya Hospital, Central South University, Changsha, China

The 15-item Geriatric Depression Scale (GDS-15) is widely used to screen depression among elders. But the factor structure of the Chinese version GDS-15 remains unclear. This study was conducted to determine the best-fit factor structure of GDS-15 and to assess measurement invariance across gender groups in a sample of Chinese elders recruited from Mainland China (final sample *N* = 2428). The best-fit factor structure was examined by confirmatory factor analysis (CFA). Multigroup CFA was utilized to test the measurement invariance across genders of the factor structure. The results of CFA revealed that a three-factor model, including life satisfaction (four items), general depressive affect (seven items), and withdrawal (three items), fits the structure of the GDS-15 best. Measurement invariance across genders was supported, fully assuming different degrees of invariance.

Keywords: depression, factor structure, measurement invariance, Chinese elders, gender differences

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Francesca Chiesi,
University of Florence, Italy
István Tóth-Király,
Concordia University, Canada

****Correspondence:***
Shuqiao Yao
shuqiaoyao@csu.edu.cn

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

***Received:** 07 March 2019*
***Accepted:** 24 May 2019*
***Published:** 21 June 2019*

Citation:
Zhao H, He J, Yi J and Yao S (2019) Factor Structure and Measurement Invariance Across Gender Groups of the 15-Item Geriatric Depression Scale Among Chinese Elders. Front. Psychol. 10:1360. doi: 10.3389/fpsyg.2019.01360

INTRODUCTION

Depression is a common mental disorder among older adults, with some 15% of community-dwelling older adults experiencing clinically significant depressive symptoms (Blazer, 2003). Late-life depression is linked to serious consequences, such as impaired daily functioning, increased health care use, and reduced quality of life (Castelo et al., 2010). Hence, assessment of depressive symptoms is an important mental health evaluation in this population.

The Geriatric Depression Scale (GDS), which was the first screening instrument to be tailored to geriatric patients (Yesavage et al., 1982), has become widely used to measure depression levels in the elderly. To reduce the time required for GDS administration and thus avoid respondent fatigue, a 15-item short-form GDS was developed from the original 30-item scale (Sheik and Yesavage, 1986). Unlike other depression tools such as the Epidemiological Studies Depression Scale (CES-D) and the Beck Depression Inventory (BDI), both versions of the GDS do not contain somatic items that may be less valid because they are common in elders (Sheik and Yesavage, 1986; Stiles and Mcgarrahan, 1998). Moreover, items of GDS use an easy response format (yes/no) preferred among older respondents. The 15-item GDS (GDS-15) retains the advantages of the original 30-item GDS, including simplicity of administration, an easy response format, and economy of time, and its validity and reliability have been demonstrated repeatedly (Cwikel and Ritchie, 1989; Lesher and Berryhill, 1994; Almeida and Almeida, 1999; Fountoulakis et al., 1999; Tang et al., 2005; Chaaya et al., 2008). Both ICD-10 criteria and DSM-IV criteria have shown that the GDS-15 is valid for measuring depression (Almeida and Almeida, 1999). GDS-15 may have more practical appeal because of the time restraints faced in clinical practice (Yao et al., 2009). In addition, the scale has been translated into multiple languages and translated versions have been proved for assessing depressive symptoms in people from various ethnic backgrounds (Iwamasa et al., 1998; Liu et al., 1998; Ishine et al., 2005; Malakouti et al., 2006; Onishi et al., 2006; Chiesi et al., 2018), including ethnic Chinese people living in Western countries (Mui, 1996; Lai, 2000).

Although the psychometric properties of the long and short GDS scales have been documented (Jang et al., 2001; Broekman et al., 2008; Pocklington et al., 2016), the factor structure of the Chinese version GDS-15 is still unclear. Mitchell et al. (1993) first proposed a three-factor model: general depressive affect (seven items), life satisfaction (four items), and withdrawal (three items). Item 10 "memory" failed to fit any of these factors. However, a number of other studies have reported different GDS-15 structures with two (Mui, 1996; Friedman et al., 2005; Brown et al., 2007), three (Incalzi et al., 2003; Imai et al., 2014), and four (Onishi et al., 2004; Lai et al., 2010) factors. Results of previous studies investigating the factor structure of the Chinese version GDS-15 have been mixed. Mui (1996) reported a two-factor model consisting of "happy mood" and "sad mood." Implementing the GDS-15 among aging Chinese in Canada, Lai and Colleagues reported a two-factor model (i.e., affective mood, cognitive mood; Lai, 2000) and a more detailed four-factor model (i.e., positive mood, negative mood, inferiority/disinterested, uncertainty, Lai et al., 2005). Most subjects of the studies above lived in Western societies. Only one study employing exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) reported a four-factor solution focused on depression among aging Chinese in Mainland China, with the following factors: positive and negative mood, energy level, inferiority, and disinterested (Lai et al., 2010). Researchers have deduced that the differences of these factor models may be related to cultural differences in the concept and expression of depression (Kim et al., 2013). For example, dominant social values of people in Western countries are individualism and personal level democratic values, whereas Chinese living in Mainland China takes more value on collectivism and at-large benefits, due to a different political and social system. These differences above in beliefs and social contexts play an important role in personal expression of affection (Mui, 2010; Kim et al., 2013).

Findings obtained depending on samples from Western societies may not necessarily be applicable to the older adults in Mainland China. The study of Lai et al. (2010) focused only on lonely elder Chinese. It is necessary for us to examine which factor structure model is more suitable for Chinese elders, for which will be helpful for developing a standardized scoring method and enable us to explore any differences across studies. In the current study, CFA was conducted to compare factor structure models that were identified in previous studies. GDS-15 total score is usually used in practice and research. However, a total score should not be used unless the covariance between the first-order factors is adequately explained by the second-order factor (Marsh and Hocevar, 1985). There are no published studies of the second-order factor of GDS-15 reported; thus, we performed a second-order factor analysis to confirm the validity of GDS-15 total scores. The trend of women having more depression problems than men was recapitulated (Nolen-Hoeksema, 2001). Tang et al. (2005) have examined the differential item functioning (DIF) of GDS-15 items, but the study was based on a sample of Hong Kong Chinese patients with pneumoconiosis. No study has tested the measurement invariance of the GDS-15 across genders in the mainland Chinese population. As related to gender, if the measurement invariance does not hold across groups, differences in observed scores may not be directly comparable (Wang et al., 2013). The true differences across groups may be mixed with the measurement bias of assessment. Exploring measurement invariance is beneficial for increasing the accuracy of depression assessments and the comparability across groups.

Hence, to develop the Chinese version of GDS-15, the first purpose of this study was to examine the best factor structure of GDS-15 in a large representative sample. A second purpose was to test the gender invariance of the GDS-15. We employed the CFA to compare the existing factor models from previous studies. Second-order CFA was performed to confirm the validity of the GDS-15 total score. Subsequently, we assessed the measurement invariance across genders of the best-fitting model.

MATERIALS AND METHODS

Sample

The inclusion criteria were as follows: age of 60–99 years old and ethnic Chinese resident of Beijing, Hunan, and Shandong province, China. The exclusion criteria were as follows: diagnosed with severe mental illness; insufficient cognitive ability to understand the questionnaire; unable to understand Mandarin and therefore unable to complete the questionnaire; cannot fill out the questionnaire due to other reasons. This study investigated the level of depression in the elderly, with 2,470 participants, and 42 failed to respond to all GDS-15 items. The final sample of 2,428 elderly Chinese volunteers included 1,141 men (47.0%) and 1,287 women (53.0%). The mean age of the men was 73.14 years [standard deviation (*SD*) = 8.07], and the mean age of the women was 71.78 years (*SD* = 7.70).

Study Design

Postgraduate psychology researchers in China were recruited and trained to do this work. Participants completed the survey in a district activity center and elderly with visual impairment or lack of formal education would get support from researchers. The study was approved by the Ethics Committee of the Second Xiangya Hospital of Central South University. Each participant gave written informed consent prior to their inclusion in the study.

Depression Symptom Assessment

The Chinese version of the GDS-15, wherein each item was a yes or no question, was used to measure depressive symptoms. The positive depression symptom response was yes for 10 items and no for 5 items, such that a point was marked for each positive symptom response. Thus, higher values indicated more depressive symptoms. As recommended by a study conducted among Chinese elders (Boey, 2000), we adopted 8 as the cutoff score. Both validity and reliability of the GDS-15 were validated satisfactory among Chinese elders in previous studies (Mui, 1996; Liu and Guo, 2008). In the current study, the scale has been confirmed to show good internal consistency (Cronbach's α = 0.873).

Statistical Analyses

Preliminary analyses were done in SPSS Version 22 (IBM, 2013), and CFA was conducted in Mplus7.4 (Muthén and Muthén, 1998). Given that the response options of items were binary (yes and no), the maximum-likelihood (ML) estimator is not adequate as it could bias the results. The robust weighted least squares with mean and variance adjustment (WLSMV) estimator was used, which could account for the binary response scaling (Finney and DiStefano, 2013; Morin et al., 2017). The whole sample was randomly divided into sample 1 (n = 1,174) and sample 2 (n = 1,254). This method of randomly assigning a larger sample into two independent samples is a common approach (Lai et al., 2010; Wang et al., 2012; He et al., 2018).

We employed CFA in sample 1 to compare competing models and determine the best-fitting factor model. A total of seven competing models were compared (**Table 1**). Models from different versions of GDS-15 were not included in the current analysis. Regular chi-square difference tests were not conducted here for the comparison of non-nested competing models. Following generally accepted practice, we used the Tucker–Lewis index (TLI), the chi-square, comparative fit index (CFI), and root mean square error of approximation (RMSEA) to evaluate the fit of each model. CFI and TLI values $\geq$0.90 indicate adequate model fit (0.95, excellent fit), while RMSEA values $\leq$0.08 and 0.06 indicate acceptable and excellent, respectively (Kline, 2010; Vrieze, 2012).

We hypothesize that there is a higher-order factor Geriatric Depression that accounts for the commonality among first-order factors. First-order CFA was conducted in sample 2 to validate the best-fitting structure of the GDS-15 confirmed in sample 1. Subsequently, second-order CFA was performed to calculate the target coefficient that could be used to decide whether the first-order factors were adequately explained by the higher-order factor. As recommended by Comrey and Lee (2013), the magnitude of the factor loadings was interpreted as follows: $\geq$0.71, excellent; 0.63–0.70, very good; 0.55–0.62, good; 0.33–0.44, fair; $\leq$0.32, poor.

Multigroup CFA was implemented in the whole sample to test gender invariance of the best-fitting model. We considered four aspects of invariance including configural invariance (Model A), metric invariance (Model B), scalar invariance (Model C), and strict invariance (Model D). Model A was used to evaluate the structure of latent variables, and the results of which served as a baseline model. Model B was tested based on the results of configural invariance with factor loading equivalence constraints imposed to ensure similarity of the observed indicators and underlying traits across gender. Model C was based on the result of the last step and in which we constrained variable intercepts equal. Model D test was conducted with factor loadings, variable intercepts, and error variance constraints equally set. As suggested by Cheung and Rensvold (2002), CFI, TLI, and RMSEA changes were employed to evaluate invariance; ΔCFI $\leq$0.01, ΔTLL $\leq$0.01, and ΔRMSEA $\leq$0.015 were considered evidence of invariance (Cheung and Rensvold, 2002; Chen, 2007).

RESULTS

Preliminary Analyses

In the whole sample, the GDS-15 total scores had a mean (SD) of 4.03 ± 3.88 for males and 4.59 ± 4.10 for females. The GDS-15 total scores range was 0–15, with women having significantly higher scores than men (t = 3.46, df = 2,426, p < 0.05). Mean (SD) GDS-15 total scores did not differ significantly (t = 0.46, df = 2,426, p > 0.05) between sample 1 (4.29 ± 3.96) and sample 2 (4.36 ± 4.04). When score $\geq$8 was used as the cutoff score, 19.9% of the participant showed significant depressive symptoms.

Factor Structure of GDS-15

As reported in **Table 2**, we obtained good fit indexes in all examined models. CFIs, TLIs, and RMSEAs were >0.95, >0.95, and <0.08, respectively. The best-fitting model was Mitchell's three-factor model (WLSMV χ^2 = 260.316, df = 74, TLI = 0.989, CFI = 0.991, RMSEA = 0.046). Next was Brown's two-factor model (WLSMV χ^2 = 438.968, df = 89, TLI = 0.980, CFI = 0.983, RMSEA = 0.058). For item 10 in Brown's model, the factor loading loaded on its latent factor was 0.116 (<0.32), a poor

TABLE 1 | Characteristics of each factor model tested.

Model	Method	Number of factors	Factor 1 items	Factor 2 items	Factor 3 items	Factor 4 items
Mitchell et al., 1993	EFA	3	3, 4, 6, 8, 12, 14, 15	1, 5, 7, 11	2, 9, 13	—
Incalzi et al., 2003	EFA	3	1, 5, 7, 11	3, 4, 8, 10, 14	2, 12, 13, 15	—
Onishi et al., 2004	EFA	4	1, 5, 7, 11, 15	2, 3, 4, 6, 8	12, 13, 14	9, 10
Friedman et al., 2005	EFA	2	2, 3, 4, 6, 8, 9, 10, 12, 13, 14, 15	1, 5, 7, 11	—	—
Brown et al., 2007	EFA	2	2, 3, 4, 6, 8, 9, 10, 12, 14, 15	1, 5, 7, 11, 13	—	—
Lai et al., 2010	EFA + CFA	4	1, 3, 4, 6, 7, 11	5, 8, 13	12, 14, 15	2, 9, 10
Imai et al., 2014	EFA + CFA	3	2, 6, 8, 9, 10, 13, 14, 15	1, 5, 7, 11	3, 4, 12	—

EFA, exploratory factor analysis; CFA, confirmatory factor analysis.

TABLE 2 | Goodness-of-fit indices of the compared models.

Model	**WLSMV χ^2**	***df***	***p***	**TLI**	**CFI**	**RMSEA (90% CI)**
Mitchell et al., 1993	260.316	74	<0.01	0.989	0.991	0.046 (0.040, 0.052)
Incalzi et al., 2003	390.318	62	<0.01	0.980	0.984	0.067 (0.061, 0.074)
Onishi et al., 2004	461.223	84	<0.01	0.977	0.982	0.062 (0.056, 0.067)
Friedman et al., 2005	464.564	84	<0.01	0.979	0.982	0.060 (0.055, 0.065)
Brown et al., 2007	438.968	89	<0.01	0.980	0.983	0.058 (0.053, 0.063)
Lai et al., 2010	454.825	84	<0.01	0.978	0.982	0.061 (0.056, 0.067)
Imai et al., 2014	456.059	87	<0.01	0.979	0.982	0.060 (0.055, 0.066)
First- and second-order CFA in sample 2						
First-order model	245.811	74	<0.01	0.991	0.993	0.043 (0.037, 0.049)
Second-order model	245.811	74	<0.01	0.991	0.993	0.043 (0.037, 0.049)

WLSMV, weighted least squares with mean and variance adjustment; df, degree of freedom; TLI, Tucker–Lewis index; CFI, comparative fit index; RMSEA, root mean square error of approximation; CI, confidence interval.

loading. Therefore, the best-fitting model for older Chinese was Mitchell's three-factor model. The results of first-order CFA in sample 2 showed that the three-factor model had an excellent fit to the data (**Table 2**). The correlations between the three factors in sample 1 ranged from 0.823 to 0.955 and those between the three factors in sample 2 ranged from 0.878 to 0.950 (see **Table 3**). All correlation coefficients were positive and statistically significant ($p < 0.001$).

Second-Order CFA

As can be seen from **Table 2**, the second-order model had the same fit indices with the first-order model (WLSMV $\chi^2 = 245.811$, $df = 74$, TLI = 0.991, CFI = 0.993, RMSEA = 0.043). Standardized factor loadings for the second-order CFA were included in **Table 4**. The first-order factor loadings ranged from 0.552 to 0.997, showing that all items were loaded well on their latent factor. The second-order factor loadings were excellent, ranging from 0.913 to 0.987 (all >0.71).

Measurement Invariance Across Genders

Given that the first-order and second-order factor model had the same fit indices, we did not test the factorial invariance of the second-order model. The results showed that the three-factor model of GDS-15 is an excellent fit of the data in both males and females. Results of multigroup CFA revealed that measurement invariance across gender groups was entirely supported at the factorial structure and the strict level (see **Table 5**). The ΔCFIs, ΔTLIs, and ΔRMSEAs are lower than 0.01 in all models, suggesting that the gender invariance of GDS-15 has been confirmed. GDS-15 items have the same meanings across genders; that is, we can compare the latent mean differences across these groups.

TABLE 3 | Factor correlations.

Factor	**Sample 1 (*n* = 1174)**			**Sample 2 (*n* = 1254)**		
	GDA	**LS**	**W**	**GDA**	**LS**	**W**
LS	0.955*			0.950*		
W	0.823*	0.885*		0.878*	0.902*	

*GDA, general depressive affect; LS, life satisfaction; W, withdrawal. *p < 0.001.*

DISCUSSION

The 15-item Geriatric Depression Scale is a widely used questionnaire for evaluating late-life depression. This study determined the best factor structure of GDS-15 suitable for Chinese elders, and it is the first to employ second-order CFA to examine the validity of the GDS-15 total score. It is also the first study to examine the factorial invariance of the GDS-15 across gender groups among Chinese elders. The findings support that the GDS-15 is a valid instrument for screening depression and as a favorable choice in situation where economy of time is required.

Several previously reported alternative best-fit models were examined by CFA. Our CFA results revealed that the best factor structure of GDS-15 suitable for Chinese elders was the original

TABLE 4 | Standardized factor loadings for the second-order CFA.

Item content	**GDA**	**LS**	**W**
3. Your life is empty	0.879		
4. Often get bored	0.849		
6. Afraid something bad will happen	0.589		
8. Often feel helpless	0.928		
12. Feel pretty worthless	0.930		
14. Situation is hopeless	0.941		
15. Most people are better off than you	0.636		
1. Satisfied with life		0.997	
5. In good spirits		0.807	
7. Happy most of the time		0.943	
11. Wonderful to be alive now		0.620	
2. Dropped activities, interests			0.558
9. Prefer to stay at home			0.552
13. Feel full of energy			0.835
Second-order factor loadings	0.962	0.987	0.913

GDA, general depressive affect; LS, life satisfaction; W, withdrawal.

TABLE 5 | Goodness-of-fit indices and model comparisons for measurement invariance models.

Model	χ^2	*df*	TLI	CFI	RMSEA (90% CI)		ΔTLI	ΔCFI	ΔRMSEA
Females	248.289	74	0.991	0.992	0.043 (0.037, 0.049)		—	—	—
Males	279.285	74	0.988	0.990	0.049 (0.043, 0.056)		—	—	—
A	921.083	148	0.935	0.947	0.066 (0.062, 0.070)		—	—	—
B	934.327	159	0.940	0.947	0.063 (0.059, 0.067)	vs. A	0.005	0.000	−0.003
C	967.045	170	0.942	0.946	0.062 (0.058, 0.066)	vs. B	0.002	−0.001	−0.001
D	1, 050.860	184	0.942	0.941	0.062 (0.059, 0.066)	vs. C	0.000	−0.005	0.000

Model A, configural invariance; Model B, metric invariance; Model C, scalar invariance; Model D, strict invariance; df, degrees of freedom; TLI, Tucker–Lewis Index; CFI, comparative fit index; RMSEA, root mean square error of approximation.

three-factor model (i.e., general depressive affect, life satisfaction, and withdrawal). Item #10 "memory problems" was dropped from the three-factor model. The factor loadings of item 10 in other models were loaded poorly on their latent factor, suggesting that the most suitable factor structure of Chinese version GDS-15 was best explained by only 14 of the 15 items. Memory problems may be attributed to the aging process. Items (1, 5, 7, and 11) of life satisfaction were common items composing one factor (Friedman et al., 2005; Brown et al., 2007; Imai et al., 2014). Items (3, 4, 6, and 8) of the first factor were also common items composing one factor (Incalzi et al., 2003; Onishi et al., 2004). These findings indicate that the symptoms of depression are at least partly consistent across diverse geriatric populations. The best factor model of GDS-15 for Chinese elders implies the three sub-dimensions in late-life depression: general depressive affect, life satisfaction, and withdrawal. It is beneficial for us to detect and prevent late-life depression from these three aspects, which will improve the efficiency of primary care. The three factors were significantly correlated with each other both in sample 1 and in sample 2, indicating that the scale has high validity. The excellent second-order factor loadings indicated that first-order factors were adequately explained by the higher-order factor. The use of GDS-15 total score was meaningful. To the best of our knowledge, this study is the first study employing second-order factor analysis to examine the validity of the GDS-15 total score. It has significant meaning for both researchers and clinicians.

In order to compare the true differences across groups, assessment tools must be measurement invariant (Wu et al., 2012). The second purpose was to evaluate the measurement invariance of depressive symptoms across genders among Chinese elders. The three-factor structure of GDS-15 was well fitted to the data in both males and females. Multiple confirmatory factors showed that measurement invariance was supported, fully assuming different degrees of invariance. The establishment of configural invariance suggests that the number of factors and factor patterns of GDS-15 is equivalent among male and female. The determination of weak equivalence indicates that the observation items and potential factors of the scale have the same meaning across groups. Satisfying strong equivalence indicates that the cross-group difference of the observed variable mean can estimate the inter-group difference of the latent variable mean. The strict equivalence, which is the most stringent equivalent based on strong equivalence, reflects cross-group differences in latent variable variation. The results of this study confirm that GDS-15 is strictly equivalent, supporting that the GDS-15 factors have the same meaning across genders. Thus, comparisons of GDS-15 scores between men and women are meaningful. It is important that studies take measurement invariance into consideration when conducting cross-group research. Together with a recent work (He et al., 2018), our study supports the notion that the GDS (both the Long and the Short form) is a reliable, valid screening instrument for detecting depression in elderly Chinese individuals, with measurement invariance across genders. Owing to its ease of administration and short period of requirement, the GDS-15 is particularly useful in situations where the economy of time is required.

Several limitations of the present work should be acknowledged. Firstly, although all the study participants were from one of three provinces in China, they were otherwise heterogeneous in terms of gender, age, economic status, education, ethnicity, and region. These undetermined sample characteristics may exist in relation to gender differences in the GDS-15. Thus, the present results generalized to other dissimilar groups remain to be determined. Secondly, because the elderly with dementia or severe physical illness were excluded from this study, the current findings may not be applicable to these groups. Thirdly, our sample consisting of older Chinese cannot represent the worldwide population. Finally, validation of the gender invariance of this Chinese version of GDS-15 does not mean that the scale has invariance across time and culture, which should be determined in future research.

CONCLUSION

In conclusion, this study found that a three-factor model fitted the underlying structure of the Chinese version of GDS-15 best. The use of GDS-15 total score is valid. In addition, the three-factor structure of GDS-15 was shown to be invariant across gender groups. Therefore, the report of significant higher GDS-15 scores of females than males reflects a true gender difference, indicating that women have more depression problems than men in aging Chinese.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The study was approved by the Ethics Committee of the Second Xiangya Hospital of Central South University.

AUTHOR CONTRIBUTIONS

All authors revised and approved the submitted version. HZ performed the initial analyses and wrote the manuscript. JH helped with collecting the data and data analysis. SY and JY supervised the study.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant No. 81871074) and the National Science and Technology Project for Professional Basic Research (Grant No. 2015FY111600).

REFERENCES

Almeida, O. P., and Almeida, S. A. (1999). Short versions of the geriatric depression scale: a study of their validity for the diagnosis of a major depressive episode according to ICD-10 and DSM-IV. *Int. J. Geriatr. Psychiatry* 14, 858–865. doi: 10.1002/(SICI)1099-1166(199910)14:10<858::AID-GPS35>3.0.CO;2-8

Blazer, D. G. (2003). Depression in late life: review and commentary. *J. Gerontol. A Biol. Sci. Med. Sci.* 58, 249–265.

Boey, K. W. (2000). The use of GDS-15 among the older adults in Beijing. *Clin. Gerontol.* 21, 49–60. doi: 10.1300/J018v21n02_05

Broekman, B. F. P., Nyunt, S. Z., Niti, M., Jin, A. Z., Ko, S. M., Kumar, R., et al. (2008). Differential item functioning of the geriatric depression scale in an asian population. *J. Affect. Disord.* 108, 285–290. doi: 10.1016/j.jad.2007.10.005

Brown, P. J., Woods, C. M., and Storandt, M. (2007). Model stability of the 15-item geriatric depression scale across cognitive impairment and severe depression. *Psychol. Aging* 22, 372–379. doi: 10.1037/0882-7974.22.2.372

Castelo, M. S., Coelho, J. M., Carvalho, A. F., Lima, J. W. O., Noleto, J. C. S., Ribeiro, K. G., et al. (2010). Validity of the Brazilian version of the geriatric depression scale (GDS) among primary care patients. *Int. Psychogeriatr.* 22, 109–113. doi: 10.1017/s1041610209991219

Chaaya, M., Sibai, A. M., El Roueiheb, Z., Chemaitelly, H., Chahine, L. M., Al-Amin, H., et al. (2008). Validation of the Arabic version of the short geriatric depression scale (GDS-15). *Int. Psychogeriatr.* 20, 571–581. doi: 10.1017/s1041610208006741

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Res. Aging* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Chiesi, F., Primi, C., Pigliautile, M., Baroni, M., Ercolani, S., Paolacci, L., et al. (2018). Does the 15-item geriatric depression scale function differently in old people with different levels of cognitive functioning? *J. Affect. Disord.* 227, 471–476. doi: 10.1016/j.jad.2017.11.045

Comrey, A. L., and Lee, H. B. (2013). *A First Course in Factor Analysis*. New York, NY: Psychology Press.

Cwikel, J., and Ritchie, K. (1989). Screening for depression among the elderly in Israel: an assessment of the short geriatric depression scale (S-GDS). *Isr. J. Med. Sci.* 25, 131–137.

Finney, S. J., and DiStefano, C. (2013). "Non-normal and categorical data in structural equation modeling," in *Structural Equation Modeling: A Second Course*. 2nd Edn, eds G. R. Hancock and R. O. Mueller (Greenwich, CO: IAP), 439–492.

Fountoulakis, K. N., Tsolaki, M., Iacovides, A., Yesavage, J., O'Hara, R., Kazis, A., et al. (1999). The validation of the short form of the geriatric depression scale (GDS) in Greece. *Aging Clin. Exp. Res.* 11, 367–372. doi: 10.1007/bf03339814

Friedman, B., Heisel, M. J., and Delavan, R. L. (2005). Psychometric properties of the 15-item geriatric depression scale in functionally impaired, cognitively intact, community-dwelling elderly primary care patients. *J. Am. Geriatr. Soc.* 53, 1570–1576. doi: 10.1111/j.1532-5415.2005.53461.x

He, J. Y., Zhong, X., and Yao, S. Q. (2018). Factor structure of the geriatric depression scale and measurement invariance across gender among chinese elders. *J. Affect. Disord.* 238, 136–141. doi: 10.1016/j.jad.2018.04.100

Imai, H., Yamanaka, G., Ishimoto, Y., Kimura, Y., Fukutomi, E., Chen, W.-L., et al. (2014). Factor structures of a Japanese version of the Geriatric Depression Scale and its correlation with the quality of life and functional ability. *Psychiatry Res.* 215, 460–465. doi: 10.1016/j.psychres.2013.12.015

Incalzi, R. A., Cesari, M., Pedone, C., and Carbonin, P. U. (2003). Construct validity of the 15-item geriatric depression scale in older medical inpatients. *J. Geriatr. Psychiatry Neurol.* 16, 23–28. doi: 10.1177/0891988702250532

Ishine, M., Wada, T., Sakagami, T., Dung, P. T., Vienh, T. D., Kawakita, T., et al. (2005). Comprehensive geriatric assessment for community-dwelling elderly in Asia compared with those in Japan: III. Phuto in Vietnam. *Geriatr. Gerontol. Int.* 5, 115–121. doi: 10.1111/j.1447-0594.2005.00277.x

Iwamasa, G. Y., Hilliard, K. M., and Kost, C. R. (1998). The geriatric depression scale and japanese american older adults. *Clin. Gerontol.* 19, 13–24. doi: 10.1300/J018v19n03_03

Jang, Y., Small, B. J., and Haley, W. E. (2001). Cross-cultural comparability of the geriatric depression scale: comparison between older koreans and older americans. *Aging Ment. Health* 5, 31–37. doi: 10.1080/13607860020020618

Kim, G., DeCoster, J., Huang, C. H., and Bryant, A. N. (2013). A meta-analysis of the factor structure of the geriatric depression scale (GDS): the effects of language. *Int. Psychogeriatr.* 25, 71–81. doi: 10.1017/s1041610212001421

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling*. 3rd Edn. New York, NY: The Guilford Press.

Lai, D., Tong, H., Zeng, Q., and Xu, W. (2010). The factor structure of a chinese geriatric depression scale-sf: use with alone elderly chinese in shanghai, china. *Int. J. Geriatr. Psychiatry* 25, 503–510. doi: 10.1002/gps.2369

Lai, D. W., Fung, T. S., and Yuen, C. T. (2005). The factor structure of a chinese version of the geriatric depression scale. *Int. J. Psychiatry Med.* 35, 137–148. doi: 10.2190/crk0-cbn7-qeve-xwpg

Lai, D. W. L. (2000). Measuring depression in Canada's elderly Chinese population: use of a community screening instrument. *Can. J. Psychiatry* 45, 279–284. doi: 10.1177/070674370004500308

Lesher, E. L., and Berryhill, J. S. (1994). Validation of the geriatric depression scale—short form among inpatients. *J. Clin. Psychol.* 50, 256–260. doi: 10.1002/1097-4679(199403)50:2<256::AID-JCLP2270500218>3.0.CO;2-E

Liu, C. Y., Lu, C. H., Yu, S., and Yang, Y. Y. (1998). Correlations between scores on chinese versions of long and short forms of the geriatric depression scale among elderly chinese. *Psychol. Rep.* 82, 211–214. doi: 10.2466/pr0.82.1.211-214

Liu, L. J., and Guo, Q. (2008). Life satisfaction in a sample of empty-nest elderly: a survey in the rural area of a mountainous county in China. *Qual. Life Res.* 17, 823–830. doi: 10.1007/s11136-008-9370-1

Malakouti, S. K., Fatollahi, P., Mirabzadeh, A., Salavati, M., and Zandi, T. (2006). Reliability, validity and factor structure of the GDS-15 in Iranian elderly. *Int. J. Geriatr. Psychiatry* 21, 588–593. doi: 10.1002/gps.1533

Marsh, H. W., and Hocevar, D. (1985). Application of con?rmatory factor analysis to the study of self-concept: first- and higher order factor models and their invariance across groups. *Psychol. Bull.* 97, 562–582. doi: 10.1037//0033-2909.97.3.562

Mitchell, J., Mathews, H. F., and Yesavage, J. A. (1993). A multidimensional examination of depression among the elderly. *Res. Aging* 15, 198–219. doi: 10.1177/0164027593152004

Morin, A. J., Arens, A. K., Tracey, D., Parker, P. D., Ciarrochi, J., Craven, R. G., et al. (2017). Self-esteem trajectories and their social determinants in adolescents with different levels of cognitive ability. *Am. J. Intellect. Dev. Disabil.* 122, 539–560. doi: 10.1352/1944-7558-122.6.539

Mui, A. C. (1996). Geriatric depression scale as a community screening instrument for elderly Chinese immigrants. *Int. Psychogeriatrics* 8, 445–458. doi: 10.1017/s1041610296002803

Mui, A. C. (2010). Productive ageing in China: a human capital perspective. *China J. Soc. Work* 3, 112–124.

Muthén, L. K., and Muthén, B. O. (1998). *Mplus User's Guide,* 7th Edn. Los Angeles, CA: Muthén and Muthén.

Nolen-Hoeksema, S. (2001). Gender differences in depression. *Curr. Dir. Psychol. Sci.* 10, 173–176. doi: 10.1111/1467-8721.00142

Onishi, J., Suzuki, Y., Umegaki, H., Endo, H., Kawamura, T., and Iguchi, A. (2006). A comparison of depressive mood of older adults in a community, nursing homes, and a geriatric hospital: factor analysis of geriatric depression scale. *J. Geriatr. Psychiatry Neurol.* 19, 26–31. doi: 10.1177/0891988705284725

Onishi, J., Umegaki, H., Suzuki, Y., Uemura, K., Kuzuya, M., and Iguchi, A. (2004). The relationship between functional disability and depressive mood in Japanese older adult inpatients. *J. Geriatr. Psychiatry Neurol.* 17, 93–98. doi: 10.1177/0891988704264738

Pocklington, C., Gilbody, S., Manea, L., and McMillan, D. (2016). The diagnostic accuracy of brief versions of the geriatric depression scale: a systematic review and meta-analysis. *Int. J. Geriatr. Psychiatry* 31, 837–857. doi: 10.1002/gps.4407

Sheik, J. L., and Yesavage, J. A. (1986). Geriatric depression scale (GDS): recent evidence and develpment of shorter version. *Clin. Gerontol.* 5, 165–173. doi: 10.1300/J018v05n01_09

Stiles, P. G., and Mcgarrahan, J. F. (1998). The geriatric depression scale: a comprehensive review. *J. Clin. Geropsychol.* 4, 89–110.

Tang, W. K., Wong, E., Chiu, H. F., Lum, C. M., and Ungvari, G. S. (2005). The geriatric depression scale should be shortened: results of rasch analysis. *Int. J. Geriatr. Psychiatry* 20, 783–789. doi: 10.1002/gps.1360

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychol. Methods* 17, 228–243. doi: 10.1037/a0027127

Wang, M. C., Armour, C., Wu, Y., Ren, F., Zhu, X. Z., and Yao, S. Q. (2013). Factor structure of the CES-D and measurement invariance across gender in Mainland Chinese adolescents. *J. Clin. Psychol.* 69, 966–979. doi: 10.1002/jclp.21978

Wang, M. C., Elhai, J. D., Dai, X. Y., and Yao, S. Q. (2012). Longitudinal invariance of posttraumatic stress disorder symptoms in adolescent earthquake survivors. *J. Anxiety Disord.* 26, 263–270. doi: 10.1016/j.janxdis.2011.12.009

Wu, W. F., Lu, Y. B., Tan, F. R., Yao, S. Q., Steca, P., Abela, J. R. Z., et al. (2012). Assessing measurement invariance of the Children's Depression Inventory in Chinese and Italian primary school student samples. *Assessment* 19, 506–516. doi: 10.1177/1073191111421286

Yao, S. Q., Zeng, H., and Sun, S. Y. (2009). Investigation on status and influential factors of cognitive function of the community-dwelling elderly in Changsha City. *Arch. Gerontol. Geriatr.* 49, 329–334. doi: 10.1016/j.archger.2008.11.00

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., et al. (1982). Development and validation of a geriatric depression screening scale: a preliminary report. *J. Psychiatr. Res.* 17, 37–49. doi: 10.1016/0022-3956(82)90033-4

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 04 July 2019
doi: 10.3389/fpsyg.2019.01440

Assessing Statistical Anxiety Among Online and Traditional Students

Marta Frey-Clark[1], *Prathiba Natesan*[1]* *and Monique O'Bryant*[2]

[1]*Educational Psychology, University of North Texas, Denton, TX, United States,* [2]*Atlanta Public Schools, Atlanta, GA, United States*

The purpose of this study was to determine whether scores on the Statistical Anxiety Scale (SAS) manifest in the same way for students in online and traditional statistics courses. Tests of measurement invariance indicated that invariance of the two-factor model of the SAS held at every level. Therefore, we compared the statistical anxiety of online and traditional students. Results indicated that online and traditional statistics students reported comparable levels of anxiety with slightly less anxiety in terms of seeking help for traditional students. We concluded that online instruction is a viable form of statistics education at least for undergraduate students enrolled in the social sciences.

Keywords: statistical anxiety, online education, measurement invariance, statistics education, validity

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Caterina Primi,
University of Florence, Italy
Thomas A. DeVaney,
Southeastern Louisiana University, United States

****Correspondence:***
Prathiba Natesan
prathiba.natesan@unt.edu

Specialty section:
This article was submitted to Quantitative Psychology and Measurement,
a section of the journal Frontiers in Psychology

Received: *17 March 2019*
Accepted: *04 June 2019*
Published: *04 July 2019*

Citation:
Frey-Clark M, Natesan P and O'Bryant M (2019) Assessing Statistical Anxiety Among Online and Traditional Students.
Front. Psychol. 10:1440.
doi: 10.3389/fpsyg.2019.01440

Participation in online education has grown rapidly over the past 15 years and is expected to continue growing (Allen and Seaman, 2010). In fact, the New York Times declared the year 2012 as the "year of the MOOC" (massive open online courses, Pappano, 2012). In Fall 2015, 29.8% of the students were enrolled online in postsecondary institutions (NCES, 2015). The online learning consortium report further shows how in addition to education, professional development, and other related sources of knowledge have moved digitally (OLC Report, 2018). Indeed, online courses seem to offer distinct advantages, with being a more convenient and cost-effective alternative to traditional, face-to-face instruction. Researchers have worked to keep pace with the growth in online learning, comparing learning outcomes for students enrolled in online courses with those of students enrolled in traditional courses.

Although several meta-analyses have shown that there was no statistically significant difference between instruction employing technology and traditional instruction (Cavanaugh et al., 2004; Zhao et al., 2005; Jahng et al., 2007), other meta-analyses have found a statistically significant difference between online and traditional instruction (Shachar and Neumann, 2003; Allen and Seaman, 2004; Bernard et al., 2004; Sitzmann et al., 2006; Williams, 2006). In fact, students with low GPAs tend to withdraw more from an online course than from a traditional course and online students tend to persist less in their programs to attain a degree (Jaggars et al., 2013). Jaggars (2014) also reported that students reported having to "teach themselves" in an online class. With respect to performance although there was a statistically significant relationship between course format (online vs. traditional) and failure in the course for English and Math courses, this was not the case for Economics and Humanities courses (Griffiths et al., 2014). Thus, it seems that there is a difference in the relationship between student performance and course format by subject matter.

Given the prevalence of anxiety in statistics courses that are perceived to be challenging, several researchers have compared performance outcomes for students enrolled in online and traditional statistics courses. Some authors have reported no difference between the two class formats (McLaren, 2004; Dotterweich and Rochelle, 2012), while one study found a difference favoring traditional instruction (Scherrer, 2011). McLaren (2004) found no statistically significant difference in the grades earned by online and traditional statistics students who completed their course; however, the researcher did find that online students demonstrated a greater tendency to drop the course or "vanish," failing to take part in assignments and exams despite remaining on the roster. Similarly, Dotterweich and Rochelle (2012) found that students enrolled in online, traditional, and televised instruction statistics courses earned similar grades; however, when the researchers isolated students who were repeating the course, they found statistically significant differences in performance favoring traditional students. By contrast, Scherrer (2011) found that when GPA, class format, and student major were included in a regression equation, class format was a statistically significant predictor of final grades, with traditional students outperforming online students.

Despite a growing body of literature comparing the performance of online and traditional statistics students, there remains a dearth of research comparing the statistical anxiety of online and traditional statistics students. Statistical anxiety is defined as "feelings of anxiety encountered when taking a statistics course or doing statistical analysis; that is, gathering, processing and interpreting data" (Cruise et al., 1985, p. 92). Statistical anxiety is a well-documented reality for statistics students (Onwuegbuzie et al., 2010; Chew and Dillon, 2014), and high statistical anxiety has consistently been associated with lower performance outcomes (Bell, 2001, 2003; Onwuegbuzie, 2004; Galli et al., 2008; Macher et al., 2012). In light of the mixed findings regarding the performance of traditional and online statistics students, as well as the documented relationship between statistics anxiety and statistics performance, it may be useful to examine the relationship between statistics anxiety and class format.

DeVaney (2010) administered a statistical anxiety pretest and posttest to traditional and online graduate students, reporting that online students had higher anxiety at the beginning of the course, but there was no difference in student anxiety at the end of the course. However, DeVaney's research operated on the assumption that measurement instrument operationalized statistical anxiety in the same way for online and traditional students. Given that previous research has identified situational antecedents to statistical anxiety (Onwuegbuzie and Wilson, 2003), it would seem that the distinct environments of traditional and online students may lead to distinct operationalization of the construct. Thus, a test of measurement invariance is a necessary foundation for future research before comparisons across traditional and online student groups can be conducted.

Measurement invariance tests the equivalence of constructs across groups along four prescribed levels (see Mellenbergh, 1989; Meredith, 1993; Vandenberg and Lance, 2000). A configural invariance model is used to test if the factor structure is defined identically across groups. Once this is established, a metric or factorial invariance model tests the equivalence of factor loadings across groups in addition to identical factor structure. Upon establishing metric invariance, a scalar invariance model is used to test if the factor structure, loadings, and item intercepts are identical across groups. Finally, an error variance invariance model is used to test if the factor structure, loadings, item intercepts, and item error variances are identical across groups. Factor means and variances may be compared only when all these levels of invariance are established. Lack of measurement invariance indicates that group-specific attributes unrelated to the latent constructs contaminate the way a person belonging to a group responds to an item (Meredith, 1993; Little, 1997). In other words, a lack of measurement invariance means that given the same factor score, individuals from different groups will have respond differently to a given item. Thus comparisons of factor scores, means, and variances in such a situation are invalid.

MEASURING STATISTICAL ANXIETY

In a review of literature on statistical anxiety, Chew and Dillon (2014) identified six extant scales, but the authors only recommended use of the Statistics Anxiety Rating Scale, or STARS (Cruise et al., 1985), and its abbreviated alternative, the Statistical Anxiety Scale, or SAS (Vigil-Colet et al., 2008). The STARS is the most widely used and well-known scale (Chew and Dillon, 2014). However, Vigil-Colet et al. (2008) criticized the STARS for its length and some of its content, which prompted their development of the SAS. The SAS has 24 items and is comprised of three subscales derived from the STARS anxiety subscales: Examination Anxiety (eight items), Interpretation Anxiety (eight items), and Asking for Help Anxiety (eight items). Examination Anxiety refers to anxiety experienced while taking a statistics test. Interpretation Anxiety refers to anxiety experienced while attempting to derive meaning from statistical formulas and output. Asking for Help Anxiety refers to anxiety experienced while requesting help of a peer, a tutor, or a professor. Each item of the SAS details a specific task, prompting respondents to indicate the level of anxiety associated with the task on a 5-point Likert-type scale ranging between *no anxiety* and *very much anxiety.*

Vigil-Colet et al. (2008) administered a Spanish version of the SAS to a sample of undergraduate students (n = 159) enrolled in statistics courses in Spain. An Exploratory Factor Analysis (EFA) verified the intended three-factor structure, with each item loading on its intended subscale. Shortly after the development and validation of the Spanish version of the SAS, Chiesi et al. (2011) administered an Italian version of the SAS to a sample of students (n = 512). A confirmatory factor analysis (CFA) confirmed the previously validated three-factor model, with the addition of correlated errors between two similarly phrased items on the Asking for Help subscale. Chiesi et al. (2011) also conducted measurement invariance tests across samples of Italian and Spanish students and reported that strict invariance of the modified three-factor model was tenable across both samples.

Following the validation of the three-factor Spanish SAS (Vigil-Colet et al., 2008) as well as the Italian SAS (Chiesi et al., 2011), O'Bryant (2017) investigated the factor structure of the English version of the SAS. After pilot-testing, she modified the items thus: Many revisions involved changing one word such as replacing *doing* to *completing* in items such as *doing a final exam in a statistics course* to *completing a final exam in a statistics course.* Other examples of changes included changing the word tutor to teacher to reflect the teaching system and terminology in the United States. O'Bryant administered the English version of the SAS to a sample of undergraduate students (n = 323) majoring in the humanities and enrolled in statistics courses throughout the United States. A CFA of the previously validated three-factor model indicated poor model fit (χ^2_{SB} = 153.46, df = 71.12, p < 0.001, RMSEA = 0.106, CFI = 0.838, SRMR = 0.073). Examination of residual correlations revealed that the residuals of the seven items on the Interpretation subscale were highly correlated with those of the items within the subscale, as well as with items on the other two subscales. Thus, O'Bryant (2017) eliminated the Interpretation subscale from the model. Eliminating the interpretation factor was not only warranted according to factor analytic output, but also seemed conceptually justifiable, given that taking an exam and asking for help are discrete tasks while interpreting numbers is not.

Further examination of residual correlations revealed that one item on the Examination Anxiety subscale and one item on the Asking for Help subscale could be eliminated due to redundancy with other items. Finally, the residuals for four items (items 1, 4, 13, and 20) on the Examination Anxiety scale were allowed to correlate, given the similarity in their wording. The resulting model had two factors, Examination Anxiety and Asking for Help Anxiety, with seven items loading on each factor and correlated errors for four items on the Examination Anxiety factor. This modified two-factor model fit the data well (χ^2_{SB} = 49.37, df = 38.13, p = 0.105, RMSEA = 0.076, CFI = 0.959, SRMR = 0.035) and was retained. We extend O'Bryant (2017) validation study to validating the factors across the online and traditional samples using measurement invariance.

The purpose of the present study is to determine whether scores on O'Bryant (2017) modified two-factor model of statistical anxiety are operationalized in the same way for traditional and online statistics students. If measurement invariance is established, an additional purpose of the present study is to compare the latent scores on the Exam Anxiety subscale and the Asking for Help Anxiety subscale for online and traditional students.

MATERIALS AND METHODS

Institutional Review Board of the University of North Texas approved the study. A two-stage sampling procedure was used. First, simple random sampling without replacement was used to randomly select institutions with social science programs to participate in the study. Second, network sampling was used to ask instructors of statistics for social science courses to pass along the research opportunity to their students. The goal was to recruit participants similar to those used in previous validation studies (Vigil-Colet et al., 2008; Chiesi et al., 2011) for comparison purposes. Data were collected online using qualtrics. Informed consent was obtained from participants who were all 18 years of age or above by asking them to click on a page that explained the study, the duration of the survey, and letting them know of the anonymity that would be maintained with the data. If they agreed to participate they could continue answering the questions by clicking on an appropriate button, else they could exit the survey. Participants were undergraduate students (n = 323) who were majoring in the social sciences and were enrolled in a statistics course. However, data screening revealed that 21 respondents took an online-traditional hybrid course, and seven respondents did not indicate their class format. Because we were only interested in online and traditional groups students, and the hybrid group was too small for analysis, these cases were dropped from the dataset, leaving 295 cases with online (n = 52) and traditional (n = 243) students. Respondents in the final dataset were predominantly female (75%), predominantly white (59%), and predominantly freshman (38%), with ages ranging from 18 to 63 years (M = 20.64, SD = 5.37).

RESULTS

Screening

The data were screened for outliers, assumptions of normality, and missing values prior to analysis. There were no outliers identified. Examination of frequency data on each item revealed severely peaked distributions, indicating that scores on the 5-point Likert-type scale were ordinal; thus, all subsequent analyses utilized non-parametric tests. Frequency data for missing values revealed a somewhat consistent distribution of missing data, with 0.3–4.7% missing per variable. Given the small percentage missing per variable and the spread of missingness across variables, data were assumed to be missing completely at random (MCAR) and were estimated *via* Mplus' default estimation for ordinal outcomes with covariates, making use of all available data to estimate missing values.

Reliability

Internal consistency of the modified two-factor SAS was measured with Cronbach's α for each class format. The α coefficients for the online class format were as follows: Total = 0.903, Exam Anxiety Subscale = 0.903, and Asking for Help Anxiety Subscale = 0.880. The α coefficients for the traditional class format were as follows: Total = 0.914, Exam Anxiety Subscale = 0.886, and Asking for Help Anxiety Subscale = 0.922. The entirety of the modified two-factor SAS and its subscales were deemed to have high internal consistent for each class format (Nunnally, 1978; Nunnally and Bernstein, 1994). McDonald's (1999) omega was computed to be 0.94 for the online class format and 0.84 for traditional class format.

Invariance Testing

We used Mplus version 7.6 with means and variance adjusted weighted least squares (WLSMV) estimation to test the measurement invariance of the SAS for online and traditional statistics students. WLSMV is a robust weighted least squares estimator that has been recommended for ordinal level data with a sample size greater than 200 (Muthén et al., 1997, unpublished; Rhemtulla et al., 2012). Because the data were ordinal, WLSMV calculates threshold parameters for each response variable to estimate the latent, continuous response indicators that correspond with each item of the SAS. Response indicators were scaled *via* theta parameterization, fixing the variance of each latent indicator to 1 in the reference group.

When comparing nested models, we used χ^2 difference tests to evaluate between-model statistical significance, with a statistically significant result indicating non-invariance across models. However, given the sensitivity of χ^2 to sample size, an *a priori* decision was made to supplement the χ^2 model testing parameters with differences in the Comparative Fit Index (CFI) and the Root Mean Square Error of Approximation (RMSEA), per Chen's (2007) criteria. Thus, the criteria for rejecting model invariance included the joint decision rules of (1) a statistically significant χ^2 difference ($p < 0.05$); (2) a change in RMSEA ≥ −0.005; and (3) a change in CFI ≤ 0.010. Note that Chen's (2007) criteria for a change in Standardized Root Mean Square Residual (SRMR) were not included because Mplus does not calculate SRMR when using WLSMV estimation to evaluate a model with covariates.

Analysis began with a confirmatory factor analysis (CFA) for each group, confirming that the O'Bryant (2017) modified two-factor model adequately fit the online group and the traditional group individually. Therefore, measurement invariance was testing by first fitting Model A that is the configural invariance model by fixing the factor structure to be identical across groups. Goodness of fit indices and approximate fit indices were tenable, indicating that the factor structure was the same for each group.

Model B, that is, the metric invariance model, was fitted by retaining the factor structure of Model A and adding constraints on all factor loadings to be equal across groups. Model fit was tenable and was not statistically significantly different from Model A, indicating that the Exam Anxiety factor and Asking for Help Anxiety factor were manifested in the same way across groups. That is, the relationships between these factors and the items that indicate them were identical across online and traditional statistics class formats. Note that the χ^2 values produced by WLSMV estimation are corrected for ordinal level data. As such, the χ^2 difference tests for nested models were also corrected by way of the DIFFTEST option in Mplus.

Model C that is, the scalar invariance was fitted by retaining constraints on factor loadings and adding constraints on item thresholds. For interval level data, testing scalar invariance would involve constraining item intercepts. However, recall that scores on items from the SAS were deemed ordinal; as such, thresholds for response options determine scores on a latent response variable, which indicates the latent factor. Thus, scalar invariance requires each threshold for each indicator to be equal across groups. Fit indices for Model C were tenable, and the fit was not appreciably worse than Model B. Therefore, the scalar invariance model was retained.

Finally, Model D, was used to test strict or error variance invariance by fixing all error variances to 1. This test deviated again from invariance testing with interval level data, in which strict invariance is established by constraining the error variances. Recall that the latent response indicators were scaled *via* theta parameterization, fixing each variance to 1 in the reference group. Thus, strict invariance was tested by fixing the latent indicator variances to 1 in both groups. Again, model fit was tenable. The scaled χ^2 difference test reported a statistically significant difference in fit compared with Model C. However, Chen's (2007) criteria for assessing differences in model fit using CFI and RMSEA did not indicate appreciably worse fit. Model D was retained, and we concluded that the SAS measures the statistical anxiety of students in online and traditional statistics classes identically. See **Table 1** for overall and comparative fit indices.

The unstandardized estimates of Model D for both groups are displayed in **Figure 1**. We note that we report unstandardized estimates because these are comparable across groups of different sample sizes. Standardized factor loadings for the online group ranged from 0.682 to 0.856; all were statistically significant at the 0.001 level. The correlation between the exam factor and help factor for the online group was 0.554, indicating the factors were related but distinct. Standardized factor loadings for the traditional group ranged from 0.659 to 0.886; again, all loadings were statistically significant at the 0.001 level. The correlation between the exam factor and help factor was 0.591, again indicating the factors were related but distinct.

TABLE 1 | Values of selected fit statistics for measurement invariance hypotheses for modified two-factor model of statistics anxiety analyzed across online and traditional student samples.

Model	Model name	χ^2_{SB}	df	Model comparison	RMSEA	CFI
Online		89.144	71		0.07 [0.0, 0.012]	0.983
Traditional		99.212			0.04 [0.018, 0.058]	0.995
Model A		185.71	142		0.046 [0.024, 0.053]	0.993
Model B	Metric invariance	198.718	154	B vs. A	0.044 [0.023, 0.061]	0.993
Model C	Scalar invariance	239.053	194	C vs. B	0.04 [0.019, 0.056]	0.993
Model D	Error variance invariance	261.041	208	D vs. C	0.042 [0.023, 0.057]	0.992

CI, confidence interval. All results were computed in Mplus for theta parameterization.

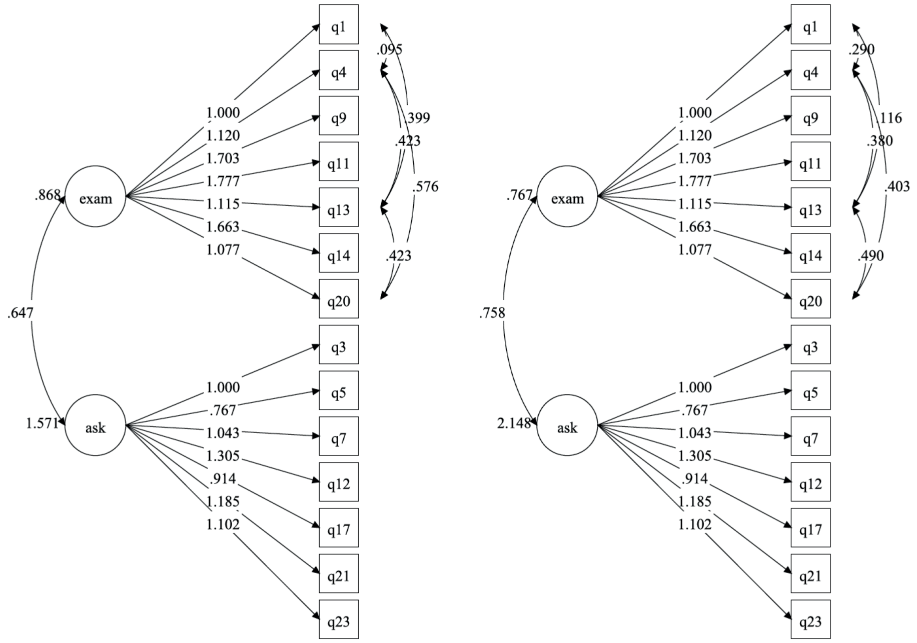


FIGURE 1 | Unstandardized estimates for traditional and online groups.

TABLE 2 | Robust weighted least squares estimates of unconstrained parameters for Model D of statistics anxiety analyzed across online and traditional student samples.

	Online			Traditional		
Parameter	**Unstd**	**SE**	**Std**	**Unstd**	**SE**	**Std**
Exam factor						
Variance	0.87	0.25	1	0.77	0.17	1
Mean	0	–	0	0.05	0.15	0.05
Question factor						
Variance	1.57	0.41	1	2.15	0.35	1
Mean	0	–	0	−0.18	0.21	−0.13
Factor covariance	0.65	0.17	0.55	0.76	0.14	0.59

Std, Standardized; Unstd, Unstandardized.

Differences in Statistical Anxieties

Having established the measurement invariance of the modified two-factor SAS for online and traditional students, analysis proceeded with the primary purpose of this study: determining by how much the two groups differed in their average scores on the Exam Anxiety subscale and the Asking for Help Anxiety subscale. See **Table 2** for the variances and means of each factor for each group. Note that the online group served as the reference group and its factor means were fixed to 0. As such, the factor means listed for the traditional group represent mean differences across groups. The mean difference in Exam Anxiety was 0.048, with online students indicating lower Exam Anxiety. The mean difference in Asking for Help Anxiety was 0.184, with online students indicating higher Asking for Help Anxiety. Cohen's d effect sizes were calculated for both mean differences, revealing effect sizes for Exam Anxiety (d = 0.054) and Asking for Help Anxiety (d = −0.129) that would be considered a very small effect (Cohen, 1988). Thus, we concluded that online statistics students expressed comparable levels of statistical exam anxiety, but slightly higher levels of asking for help anxiety than traditional statistics students.

DISCUSSION

The purpose of the present study was to determine whether the operationalization of statistical anxiety *via* the modified two-factor Statistical Anxiety Scale is the same for samples of online students and traditional students. Previous research has indicated that online statistics students may represent a distinct demographic, being older, with more credit hours earned and more courses repeated than their traditional counterparts (Dotterweich and Rochelle, 2012). Previous research has also indicated online students may possess different intellectual strengths, having higher logical-mathematical intelligence than their traditional counterparts (Lopez and Patron, 2012). If the two populations differ with respect to demographic characteristics and intellectual strengths, it may seem probable that they could differ with respect to the manner in which they report statistical anxiety. However, this was not the case.

Invariance held at every level, indicating that the modified two-factor SAS measures statistical anxiety manifests in the same way for online and traditional statistics students. These findings are further strengthened by the fact that the sample for the present study was drawn *via* random cluster sampling of colleges and universities throughout the United States. Thus, the SAS would appear to be a versatile measure of statistical anxiety. This finding answers Chew and Dillon's (2014) call to confirm the factor structure of the SAS with diverse samples and provides a foundation for future research using the SAS with classes of varied formats.

Given that the modified two-factor model of the SAS is comprised of only 14 items, and scores on these items are valid for both online and traditional students, statistics instructors may consider administering this instrument to students in order to gauge anxiety and adjust instruction accordingly. Researchers have identified a number of effective interventions, including the use of humor (Pan and Tang, 2004), problem-solving games (D'Andrea and Waters, 2002), and instructor immediacy (Williams, 2006). Thus, the SAS could serve as a diagnostic tool, presenting instructors with student feedback to inform instruction.

An added purpose of this study was to compare mean scores for Exam Anxiety and Asking for Help Anxiety across class formats. Effect size estimates revealed that mean differences were negligible for exam anxiety and a lower asking for help anxiety for traditional students. This is contrary to popular belief that students have lesser inhibitions in reaching out for help when they are learning within the relative privacy and social safety of online education. However, the effect size is too small to make conclusions regarding these differences.

Our findings lend additional support to DeVaney's (2010) finding that online and traditional students had comparable levels of anxiety upon completion of an introductory statistics course. Furthermore, DeVaney reported that online students had higher statistical anxiety than traditional students at the beginning of the course. Thus, if online students do not appear to carry greater statistical anxiety, as our study suggests, and if the online class format may even soothe statistical anxiety, as DeVaney's work suggests, then online statistics education seems to present a viable alternative to traditional, face-to-face instruction.

Institutions of higher learning have reported offering online courses in the interest of meeting student demand for flexible scheduling, providing college access to students who may not otherwise have access, making courses more available, and seeking to increase student enrollment (Parsad and Lewis, 2008). As a convenient class format for students, and a cost-effective class format for institutions of higher learning, capitalizing on the pragmatic advantages of online education may allow a greater number of students to access statistics education, and a greater number of institutions to offer statistics education.

A major limitation of the present study is its small sample size. It is recommended that this study be repeated for larger samples so as to address the generalizability of the study. Perhaps administering a pre- and post-survey to examine statistics anxiety before and after taking traditional and online courses is another avenue for future research. Future research might seek to clarify the relationship between class format, statistical anxiety, and performance outcomes. Given the established relationship between statistical anxiety and performance outcomes (e.g., Galli et al., 2008), and the conflicting findings regarding the relationship of class format to performance outcomes (e.g., Scherrer, 2011; Dotterweich and Rochelle, 2012), there exists the possibility that class format and statistical anxiety interact to influence performance outcomes. Examination of all three variables in context may serve to clarify their relationships and inform future instruction. Regardless, insofar as the present study stands, online and traditional statistics students experience similar levels of anxiety, indicating that online instruction is a viable means of delivering statistics education.

DATA AVAILABILITY

The datasets for this manuscript are not publicly available because the dataset is part of the MOB's thesis. Covariance matrix may be provided upon request. But the data are subject to confidentiality agreement according to informed consent. Requests to access the datasets should be directed to monique_obryant@yahoo.com.

ETHICS STATEMENT

The institutional review board of the university of North Texas approved this study. Informed consent was obtained from participants before they answered the survey. Vulnerable populations were not involved.

AUTHOR CONTRIBUTIONS

MF-C conducted the data analysis and literature review. PN oversaw the project and added conclusion and introduction. MOB collected the data, came up with the instrument, and helped with literature review.

REFERENCES

Allen, I. E., and Seaman, J. (2010). *Class differences: Online education in the United States*. (Newburyport, MA: Sloan Consortium).

Allen, I. E., and Seaman, J. (2004). *Entering the mainstream: the quality and extent of online education in the United States, 2003 and 2004*. Newburyport, MA: Sloan Consortium.

Bell, J. A. (2001). Length of course and levels of statistics anxiety. *Education* 121, 713–716.

Bell, J. A. (2003). Statistics anxiety: the nontraditional student. *Education* 124, 157–162.

Bernard, R., Brauer, A., Abrami, P., and Surkes, M. (2004). The development of a questionnaire for predicting online learning achievement. *Distance Educ.* 25, 31–47. doi: 10.1080/0158791042000212440

Cavanaugh, C., Gillan, K. J., Kromrey, J., Hess, M., and Blomeyer, R. (2004). *The effects of distance education on K-12 student outcomes: A meta-analysis*. Naperville, IL: Learning Point Associates/North Central Regional Educational Laboratory.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Chew, P. K. H., and Dillon, D. B. (2014). Statistics anxiety update: refining the construct and recommendations for a new research agenda. *Perspect. Psychol. Sci.* 9, 196–208. doi: 10.1177/1745691613518077

Chiesi, F., Primi, C., and Carmona, J. (2011). Measuring statistics anxiety: cross-country validity of the statistical anxiety scale (SAS). *J. Psychoeduc. Assess.* 29, 559–569. doi: 10.1177/0734282911404985

Cohen, J. (1988). *Statistical power analysis for behavioral sciences.* 2nd edn. (Hillsdale, NJ: Lawrence Earlbaum Associates).

Cruise, R. J., Cash, R. W., and Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *Proceedings of the American Statistical Association, Section on Statistical Education, Las Vegas, NV*.

D'Andrea, L., and Waters, C. (2002). Teaching statistics using short stories: reducing anxiety and changing attitudes. In: Sixth International Conference on Teaching Statistics, Cape Town, South Africa.

DeVaney, T. A. (2010). Anxiety and attitude of graduate students in on-campus vs. online statistics courses. *J. Stat. Educ.* 18. doi: 10.1080/10691898.2010.11889472

Dotterweich, D. P., and Rochelle, C. F. (2012). Online, instructional television, and traditional delivery: student characteristics and success factors in business statistics. *Am. J. Bus. Educ.* 5, 129–138. doi: 10.19030/ajbe.v5i2.6815

Galli, S., Ciancaleoni, M., Chiesi, F., and Primi, C. (2008). Who failed the introductory statistics examination? A study on a sample of psychology students. Paper presented at the 11th International Congress on Mathematical Education, Monterrey, Mexico.

Griffiths, R., Chingos, M., Mulhern, C., and Spies, R. (2014). *Interactive online learning on campus: Testing MOOCs and other hybrid formats in the University System of Maryland*. New York: Ithaka S+R.

Jaggars, S. S. (2014). Choosing between online and face-to-face courses: community college student voices. *Am. J. Dist. Educ.* 28, 23–28. doi: 10.1080/08923647.2014.867697

Jaggars, S. S., Edgecombe, N., and Stacey, G. W. (2013). *What we know about online course outcomes*. NY: Community College Research Center.

Jahng, N., Krug, D., and Zhang, Z. (2007). Student achievement in online distance education compared to face-to-face education. *Eur. J. Open Dist. Online Learn.* 10. http://www.eurodl.org/materials/contrib/2007/Jahng_Krug_Zhang.htm

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues. *Multivar. Behav. Res.* 32, 53–76. doi: 10.1207/s15327906mbr3201_3

Lopez, S., and Patron, H. (2012). Multiple intelligences in online, hybrid, and traditional business statistics courses. *J. Edu. Online* 9. doi: 10.9743/JEO.2012.2.2

Macher, D., Paechter, M., Papousek, I., and Ruggeri, K. (2012). Statistics anxiety, trait anxiety, learning behavior, and academic performance. *Eur. J. Psychol. Educ.* 27, 483–498. doi: 10.1007/s10212-011-0090-5

McDonald, R. P. (1999). *Test theory: A unified treatment*. (Mahwah, NJ: Erlbaum Associates).

McLaren, C. H. (2004). A comparison of student persistence and performance in online and classroom business statistics experiences. *Decis. Sci. J. Innov. Educ.* 2, 1–10. doi: 10.1111/j.0011-7315.2004.00015.x

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

NCES. (2015). Available at: https://nces.ed.gov/fastfacts/display.asp?id=80 (Accessed August 15, 2018).

Nunnally, J. C. (1978). *Psychometric theory.* 2nd edn. (NY: McGraw-Hill).

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric theory.* 3rd edn. (NY: McGraw-Hill).

O'Bryant, M. J. (2017). How attitudes towards statistics courses and the field of statistics predicts statistics anxiety among undergraduate social science majors: a validation of the Statistical Anxiety Scale. ProQuest LLC. Doctoral dissertation, University of North Texas. Available online at https://search.proquest.com/docview/2009455494

OLC Report (2018). Available at: https://olc-wordpress-assets.s3.amazonaws.com/uploads/2019/04/OLC-2018-Annual-Report-Online.pdf (Accessed August 15, 2018).

Onwuegbuzie, A. J. (2004). Academic procrastination and statistics anxiety. *Assess. Eval. Higher Educ.* 29, 3–19. doi: 10.1080/0260293042000160384

Onwuegbuzie, A. J., Leech, N. L., Murtonen, M., and Tähtinen, J. (2010). Utilizing mixed methods in teaching environments to reduce statistics anxiety. *Int. J. Multiple Res. App.* 4, 28–39. doi: 10.5172/mra.2010.4.1.028

Onwuegbuzie, A. J., and Wilson, V. S. (2003). Statistics anxiety: nature, etiology, antecedents, effects, and treatments—a comprehensive review of the literature. *Teach. High. Educ.* 8, 195–209. doi: 10.1080/1356251032000052447

Pan, W., and Tang, M. (2004). Examining the effectiveness of innovative instructional methods on reducing statistics anxiety for graduate students in the social sciences. *J. Instructional Psychol.* 31, 149–159.

Pappano, L. (2012). *The year of the MOOC*. https://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html

Parsad, B., and Lewis, L. (2008). *Distance education at degree-granting postsecondary institutions: 2006–2007. First look (NCES 2009–044)*. (Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics).

Rhemtulla, M., Brosseau-Liard, P. E., and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychol. Methods* 17, 354–373. doi: 10.1037/a0029315

Scherrer, C. R. (2011). Comparison of an introductory level undergraduate statistics course taught with traditional, hybrid, and online delivery methods. *INFOMRS Trans. Educ.* 11, 106–110. doi: 10.1287/ited.1110.0063

Shachar, M., and Neumann, Y. (2003). Differences between traditional and distance education academic performances: a meta-analytic approach. *Int. Rev. Res. Open Dist. Learn.* 4, 1–20. doi: 10.19173/irrodl.v4i2.153

Sitzmann, T., Kraiger, K., Steward, D., and Wisher, R. (2006). The comparative effectiveness of web-based and classroom instruction: a meta-analysis. *Pers. Psychol.* 59, 623–664. doi: 10.1111/j.1744-6570.2006.00049.x

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Vigil-Colet, A., Lorenzo-Seva, U., and Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema* 20, 174–180. doi: 10.1037/t62688-000

Williams, S. L. (2006). The effectiveness of distance education in allied health science programs: a meta-analysis of outcomes. *Am. J. Dist. Educ.* 20, 127–141. doi: 10.1207/s15389286ajde2003_2

Zhao, H., Seibert, S. E., and Hills, G. E. (2005). The mediating role of self-efficacy in the development of entrepreneurial intentions. *J. Appl. Psychol.* 90, 1265–1272. doi: 10.1037/0021-9010.90.6.1265

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 16 July 2019
doi: 10.3389/fpsyg.2019.01646

Regression Analysis of ICT Impact Factors on Early Adolescents' Reading Proficiency in Five High-Performing Countries

*Ya Xiao, Yang Liu and Jie Hu**

Department of Linguistics and Translation, School of International Studies, Zhejiang University, Hangzhou, China

The popularity of information and communication technology (ICT) has had a significant influence on the reading proficiency of early adolescents. Achieving excellent reading proficiency, which is related not only to a student's inherent talent but also to various impact factors, can greatly enhance the effectiveness of reading education. The Program for International Student Assessment (PISA) 2015 provides an international view on the reading proficiency of 15-year-olds in a computer-based testing environment. In this study, a multiple linear regression model was constructed using the computing language R to investigate the association between student-level ICT impact factors (the availability of ICT, the use of ICT and attitudes toward ICT) and reading proficiency among early adolescents. The sample included 37,155 15-year-olds from five representative countries with extremely high reading proficiency. The results showed that the students' ICT-related attitudinal factors concerning their interest in ICT and perceived autonomy in using ICT, rather than ICT availability and ICT use, were closely associated with high reading proficiency. In addition, ICT devices should be integrated not only as instructional media but also as a cognitive tool for teaching reading with timely and appropriate scrutiny.

Keywords: ICT impact factors, reading proficiency, multiple linear regression, early adolescent, PISA 2015

Edited by:
Elisa Pedroli,
Italian Institute for Auxology (IRCCS),
Italy

Reviewed by:
Paula Fariña,
Diego Portales University, Chile
Gareth J. Williams,
Nottingham Trent University,
United Kingdom

****Correspondence:***
Jie Hu
huj@zju.edu.cn
orcid.org/0000-0003-2219-2587

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *01 March 2019*
Accepted: *28 June 2019*
Published: *16 July 2019*

Citation:
Xiao Y, Liu Y and Hu J (2019) Regression Analysis of ICT Impact Factors on Early Adolescents' Reading Proficiency in Five High-Performing Countries. Front. Psychol. 10:1646. doi: 10.3389/fpsyg.2019.01646

INTRODUCTION

The concept of computer-based assessment of reading proficiency is of fundamental significance in the age of information and communication technology (ICT) (Naumann, 2015). The proliferation of ICT has a profound influence on the concept of reading proficiency (e.g., Liu, 2005; Coiro and Dobler, 2007) because it has largely reshaped students' learning processes and reading activities (e.g., Gan et al., 2015; Mantoro et al., 2017) by engaging students in effective reading activities (e.g., Chen and Hu, 2018) and improving their reading comprehension ability (e.g., Whyte et al., 2014). As the benchmark of international large-scale assessment, the Program for International Student Assessment (PISA) has evaluated reading, science, and mathematics achievement among 15-year-olds from participating countries/economies of the Organization of Economic and Cultural Development (OECD) every 3 years since 2000. Reading proficiency in this influential assessment is recognized as "students' ability to understand, use, reflect on and engage with written texts in order to achieve one's goals, develop one's knowledge and potential, and participate in society"

(OECD, 2015, p. 30). This large-scale assessment facilitates the infrastructural and epistemological construction of global education work (Sellar and Lingard, 2013). For the first time, the PISA 2015 delivered the assessments of all three subjects via computer. Among the 72 participating economies, only 15 economies took the paper-based test due to technical problems. These changes have launched a new area of research, that is, the role played by myriad ICT impact factors in students' reading proficiency because different types of reading activities and related impact factors have emerged (OECD, 2011).

The PISA reading proficiency test has been studied for nearly 20 years. From the long-term perspective, from the PISA 2000 to the PISA 2015, there has been no significant change in the framework of reading assessment among the six consecutive cycles of PISA 2000, PISA 2003, PISA 2006, PISA 2009, PISA 2012, and PISA 2015 (OECD, 2012, 2017). Thus, the whole reading framework and a large number of derived variables in the PISA 2015 were also taken from the previous PISA cycles without change as part of the trend content. In this sixth cycle of PISA assessment, a set of tasks including 103 questions was used in the PISA 2015 reading assessment (OECD, 2016, p. 146). Students' reading proficiency scores were analyzed based on item response theory and officially released in the *PISA 2015 Results*. The proficiency levels described from the lowest to the highest are Level 1b, Level 1a, Level 2, Level 3, Level 4, Level 5, and Level 6. These seven proficiency levels used in the PISA 2015 reading assessment are the same as those established for the PISA 2009 assessment. The required reading skills at each proficiency level are described according to the three processes by which students answer the questions. These three processes are defined in the framework as "access and retrieve" (skills associated with finding, selecting and collecting information); "integrate and interpret" (processing what is read to make sense of a text); and "reflect and evaluate" (drawing on knowledge, ideas or values external to the text) (OECD, 2016, p. 162).

Starting with the PISA 2009, the OECD, for the first time, designed a computer-based reading assessment as an additional option for its reading proficiency test. Regarding the assessment contents of the paper-based and computer-based PISA 2015 reading proficiency assessment, the latter differs from the former only in format, i.e., the way of presenting long texts by screen and the basic knowledge of hardware usage. However, compared with other cycles of the ICT familiarity questionnaire in the PISA computer-based assessment of reading, four derived variables were newly developed in the PISA 2015, including students' ICT interest (INTICT), perceived competence in ICT usage (COMPICT), perceived autonomy related to ICT usage (AUTICT) and the degree to which ICT is part of their daily social life (SOIAICT). In particular, the index for ICT use outside of school for academic purposes has changed over time: in the PISA 2006 ICT familiarity questionnaire, this index includes five questions that mainly address students' degree of using a computer to write papers, create spreadsheets, draw or use graphics programs, use educational software and write computer programs (OECD, 2009). In the PISA 2012, this index is examined using seven measurements of browsing the Internet for schoolwork, using email for communication with other students about schoolwork, using email for communication with teachers and the submission of homework, downloading, uploading or browsing material from the school's website, checking the school's website for announcements, doing homework on the computer, and sharing school-related materials with other students (OECD, 2015). Finally, in the PISA 2015, the index is derived from 12 measurements, including all seven measurements that were examined in the PISA 2012. In addition, students' degrees of browsing the Internet to follow up lessons, using social networks for communication with other students and teachers about schoolwork, doing homework on a mobile device, and downloading learning apps on a mobile device are also included (OECD, 2017).

Substantial effort has been made to investigate the impacts of certain factors on students' reading proficiency based on the PISA assessment framework. The previous studies can be divided into three categories. The first category is that of sociodemographic factors. Gender, family background and immigration background are confirmed to be significant sociodemographic factors of computer-based assessment measuring reading proficiency. Specifically, 15-year-old girls tend to score higher in computer-based reading assessments on multiple layers of reading skills than boys of the same age (e.g., Stoet and Geary, 2015; Puteh et al., 2016; Torppa et al., 2018). In addition, parental education (e.g., Rajchert et al., 2014), early parental engagement in educational activities (e.g., Hemmerechts et al., 2016), and parental involvement in social and cultural exchange (e.g., Gotoh et al., 2013) are found to be positive factors of reading performance. For immigrant background factors, immigrant students perform consistently worse than native students (e.g., Liberto, 2014), which can be explained by insufficient family support and the control of immigrants (Santos et al., 2016). In the meantime, it has also been found that the sense of school belonging exerts a moderating effect in the mathematical achievement gap between immigrants and natives (Schachner et al., 2017); however, for reading performance, this moderating effect turns out to be insignificant (Mok et al., 2016). The second category is related to cognitive factors. Cognitive skills (rapid naming, phonological awareness, and letter knowledge) and cognitive learning strategies (elaboration and memorization) positively influence reading proficiency (e.g., Li and Chun, 2012; Eklund et al., 2018). The third category concerns instructional factors. Categorical instructions or curricula targeting students of different reading levels improve their reading results (e.g., Shin et al., 2013). In addition, teachers' guidance of students when they encounter difficulties in reading, teachers' stimulation of students' reading processes and the classroom reading environment are all meaningful factors influencing students' reading proficiency (Meng et al., 2017).

Previous studies have constructed statistical models using the theoretically-based rationale that ICT impact factors are related to reading proficiency. For instance, to examine the mediation effect from individual differences in the inner and outer states of ICT to the PISA reading proficiency, a partial mediation model was constructed (Lee and Wu, 2012). An ordered logit model was employed to estimate relationships between an ordinal dependent variable (i.e., PISA test score) and a set of independent variables (i.e., the student's background, school characteristics, the home/family environment and the student's access to ICT facilities) (Erdogdu and Erdogdu, 2015). In these studies, special attention was given to student-level ICT impact factors. Student-level ICT impact factors were obtained from the ICT familiarity questionnaire, which has been gradually developed since the PISA 2000. In the PISA 2015 questionnaire, these factors can be generalized into three main categories: the availability of ICT, the use of ICT and attitudes toward ICT. With regard to the impact of ICT availability, the mere availability of ICT at home is negatively related to reading proficiency, whereas ICT availability at school is not significantly correlated with reading performance (e.g., Lee and Wu, 2012; Hu et al., 2018).

Two relevant contextual factors have been identified in the previous literature with a focus on the impact of ICT use on reading proficiency. The first involves where the ICT is used, i.e., at school or outside of school. The findings regarding ICT use at school are complex; the association between ICT use at school and students' reading achievement is recognized as having an inverted U-shape, which indicates that overuse of ICT at school may reverse the positive correlation between ICT use at school and students' reading proficiency (Woessmann and Fuchs, 2005); however, ICT use at school is also found to be negatively correlated with students' reading proficiency (Petko et al., 2017). Furthermore, this relationship varies among students in different grades. In particular, ICT use at school is found to be positively associated with the reading performance of fourth-grade students whereas it is negatively correlated with that of eighth-grade students (Skryabin et al., 2015). With regard to the second contextual factor, ICT is used outside of school for social entertainment or for web navigation. Specifically, the dimension of social entertainment involves the accessing of email, collaborative gaming, and the use of social media. The dimension of information seeking on the Internet includes reading online news, using e-dictionaries, consulting online encyclopedias and browsing websites for practical information. Some researchers have found that online navigation activities outside of school improve students' reading proficiency whereas leisure activities decrease it (e.g., Woessmann and Fuchs, 2005; Lee and Wu, 2013). In contrast, some scholars discover that ICT use for entertainment at home is positively correlated with students' reading performance (Skryabin et al., 2015). Additionally, ICT use for leisure is found to narrow the gender gap in students' reading scores (e.g., Cheung et al., 2013; Rasmusson and Åberg-Bengtsson, 2015).

Attitude is a significant psychological construct that inheres in or characterizes a person (Richard, 2016). With regard to the ICT attitudinal variables included in the ICT familiarity questionnaire of the PISA 2015, students' attitudes were found to positively influence students' reading performance (Lee and Wu, 2012; Petko et al., 2017). In contrast, attitudes toward ICT for social interaction are negatively associated with reading proficiency (Hu et al., 2018). Researchers have used different indexes of ICT attitudes based on the PISA ICT familiarity questionnaire that they selected. For instance, Lee and Wu (2012) obtained one attitudinal index derived from four indicators based on the PISA 2009 ICT familiarity questionnaire. Petko et al. (2017) applied positive attitude toward ICT as a learning tool (ICTATTPOS) derived from six indicators based on the questionnaire in the PISA 2012. Considering that the constructs of ICT attitudes applied in the previous studies are not yet fully developed, a more comprehensive ICT familiarity questionnaire of the PISA 2015 is utilized in the current study to analyze the impacts of students' ICT-related attitudes; this questionnaire includes four explicit indexes: interest in ICT, perceived ICT competence, perceived autonomy in using ICT, and enjoyment of social communication using ICT (OECD, 2017).

Achieving excellence in education can greatly enhance the effectiveness of education (OECD, 2009); excellence involves more than a student's inherent talent as it is also related to various interactive factors (Hu and Wei, 2018). Most of the abovementioned studies investigated the ICT impact factors of students' reading proficiency in one or more countries; however, the literature on the representativeness of countries with excellent reading proficiency remains insufficient. The top-performing countries should receive particular attention since relevant findings would certainly offer innovative insights leading to educational excellence for educators and policymakers around the world (Jerrim, 2015). Certain previous studies have investigated the relationship between impact factors and excellent subject performance by students. For instance, pedagogical impact factors of 4th-grade students with excellent reading proficiency were identified based on Progress in International Reading Literacy Study (PIRLS) (Xiao and Hu, 2019). Regarding the PISA-based analysis, a set of impact factors influencing top students' science performance was explored (e.g., Chen et al., 2019). However, few of these studies have targeted ICT impact factors and 15-year-olds' reading proficiency in high-performing countries. Therefore, this study aimed to identify the correlation between ICT impact factors and early adolescents' reading proficiency in high-performing countries based on the large-scale educational assessment of the PISA 2015. Although the examination of the high-performing countries versus the low-performing ones can maximize the research scope, such comparisons may lead to invalid conclusions and weak representations of educational success because of the polar socioeconomic situations in different countries (OECD, 2016). Therefore, the study's research objective is to survey the impact of ICT factors on secondary school students' reading performance in five representative countries with extremely high reading proficiency.

MATERIALS AND METHODS

Sample

The sample was drawn from the PISA 2015 dataset[1], which is the latest PISA dataset, released in December of 2017. Different from the previous cycles, the assessments of all three domains of science, reading and mathematics were mainly conducted on computers in the PISA 2015. Of the 72 countries/economies that participated in this international

[1] http://www.oecd.org/pisa/data/2015database/

assessment, 57 countries/economies (including all 35 OECD members) completed the computer-based assessment (CBA) whereas the remaining 15 participants who lacked computer-test access used the paper-based alternatives. Questionnaires were administered to students, principals, teachers, and parents to obtain relevant contextual information. Only the CBA countries/economies could choose whether to take the ICT familiarity questionnaire (OECD, 2017).

In the case of PISA, students are categorized into seven proficiency levels for each domain based on their test scores: Level 1b is the lowest described level, then Level 1a, Level 2, Level 3 and so on up to Level 6 as the highest proficiency level. Students reaching Level 5 or 6 on the reading proficiency scale are referred to as top performers. Level 6 tasks are more challenging and rigorous than Level 5 tasks. Students reaching Level 6 are typically able to integrate information from multiple texts, understand connotations on a sophisticated level, and expertly handle unfamiliar ideas.

According to the statistical results of the PISA 2015, only countries with at least 2% of performers at Level 6 may be regarded as representative countries with excellent reading proficiency (OECD, 2016) because high-performing educational systems can present better teaching resources, stronger school leadership, higher academic standards, broader educational outcomes, more innovative educational reforms and more international vision than others (Deng and Gopinathan, 2016). Among the seven representative countries with excellent reading proficiency, Canada and Norway did not take the ICT familiarity questionnaire. Thus, in the current study, Singapore (3.600% of Level 6 performers), New Zealand (2.600% of Level 6 performers), Australia (2.000% of Level 6 performers), Finland (2.000% of Level 6 performers) and France (2.000% of Level 6 performers) were selected as the five sample countries across Asia, Europe, and Oceania. Considering the representativeness of these five countries, all students were taken into consideration without distinguishing high- from low-achieving performers. The data of 37,155 sample students were retrieved by Perl computing language version 5.28.2. Boys account for 49.433% of the sample, and girls account for 50.567% of the sample. The age range of the participants was between 15 years and 3 (complete) months and 16 years and 2 (complete) months, as strictly required by the PISA (OECD, 2016, p. 210). In addition, the percentage of individuals with ICT availability at home (ICTHOME) or at school (ICTSCH) is at least 98.640% in five countries, respectively. Students with access to ICT both at home and at school are shown as 99.890% in total. The demographic information is presented in **Table 1**.

Data Analysis

Variables

As it is impossible for each student to complete all test items, the PISA 2015 computed 10 plausible values (PVs) of reading scores to measure students' performance (see **Table 2**). The present study followed the recommendations for addressing PVs in international large-scale assessments (OECD, 2009; Rutkowski et al., 2010), considering all 10 PVs simultaneously as the dependent variables for the purpose of obtaining unbiased and stable estimates.

This study included three categories of student-level ICT factors as regressors (see **Table 3**), i.e., the availability of ICT (at school and outside of school), the use of ICT (at school or outside of school for academic and leisure purposes), and attitudes toward ICT (students' interest in ICT, perceived autonomy related to ICT, perceived ICT competence, and ICT use for social interaction). In addition, the binary variable Gender and the derived variable of students' gender and economic, social and cultural status (ESCS) were also considered. Based on the theoretical rationale in this study, all variables related to ICT availability, ICT use and attitudes toward ICT were included in the following analyses.

Multiple Linear Regression (MLR) Modeling

A regression model that contains more than one regressor variable is called a multiple regression model (Montgomery and Runger, 2007). An MLR model is "typically employed to measure the effects of the explanatory variables on performance" (Fariña et al., 2015, p. 179). It can accurately reflect the correlations among factors, indicate the degree of fit, and improve the effect of the regression equation (Holmes and Rinaman, 2015). Linear relationships among the various factors can be analyzed intuitively and promptly by using multiple sets of data.

In this study, considering that students' reading proficiency is associated with multiple factors, it is effective and realistic to estimate the dependent variable by using the optimal combination of multiple independent variables, which can be accurately realized by an MLR model, in line with recommendations for PISA data analysis (Rutkowski et al., 2010). The equation for MLR is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon \tag{1}$$

where

y_i refers to the dependent variables,

β_0 refers to the intercept, and

β_p refers to the partial regression coefficient, which gauges the unit change in the dependent variable per unit increase in the factors on the condition that the rest of the factors remain unchanged.

ε refers to the error term.

In the current study, MLR modeling was performed using R computing language version 3.5.0[2]. The data analysis procedure was as follows:

First, the data preprocessing procedure was conducted. Large-scale assessments (e.g., the PISA), conducted in the context of item response theory (Cui et al., 2019), generally contain missing values. In this context, the aggr() function from the R Language package 'VIM' was used to visualize the number and proportion of missing values. Deleting the missing values is one solution when the missing rate is lower than 5% for each variable; however, this solution could not be used in this study due to the high missing rate of over 10%. Therefore, to ensure the maximum number of observations, the imputation

[2] https://www.r-project.org/

TABLE 1 | Demographic information of participants from five representative countries.

Country	Observation	Gender	Participants with ICT availability at school	Participants with ICT availability at home	Participants with ICT availability at school and at home
		% female	% yes	% yes	% yes
Australia	14,530	49.298% (7,163/14,530)	99.484% (14,455/14,530)	99.780% (14,498/14,530)	99.876% (14,512/14,530)
Finland	5,882	48.674% (2,863/5,882)	98.640% (5,802/5,882)	99.898% (5,876/5,882)	99.915% (5,877/5,882)
France	6,108	50.933% (3,111/6,108)	98.838% (6,037/6,108)	99.853% (6,099/6,108)	99.935% (6,104/6,108)
New Zealand	4,520	49.934% (2,257/4,520)	99.535% (4,499/4,520)	99.823% (4,512/4,520)	99.978% (4,519/4,520)
Singapore	6,115	48.618% (2,973/6,115)	98.692% (6,035/6,115)	99.920% (6,110/6,115)	99.787% (6,102/6,115)
Total	37,155	49.433% (18,367/37,155)	99.120% (36,828/37,155)	99.812% (37,085/37,155)	99.890% (37,114/37,155)

Sources: OECD PISA 2015 general database.

TABLE 2 | Descriptive statistics of plausible values of reading proficiency in the PISA 2015 computer-based reading assessment.

Variable	Description	Mean	*SD*	Minimum	Maximum
PV1READ	Plausible value 1 of reading score	506.691	104.000	73.377	851.085
PV2READ	Plausible value 2 of reading score	507.148	104.337	46.927	844.637
PV3READ	Plausible value 3 of reading score	507.810	104.006	57.679	851.970
PV4READ	Plausible value 4 of reading score	506.642	104.387	83.508	839.131
PV5READ	Plausible value 5 of reading score	507.818	104.958	96.893	865.085
PV6READ	Plausible value 6 of reading score	507.261	103.858	0.000	870.747
PV7READ	Plausible value 7 of reading score	507.667	104.540	28.659	898.018
PV8READ	Plausible value 8 of reading score	507.173	104.202	46.421	849.645
PV9READ	Plausible value 9 of reading score	508.171	104.693	22.847	864.958
PV10READ	Plausible value 10 of reading score	507.018	104.935	81.639	884.906

Sources: OECD PISA 2015 general database. N = 37,155. The dependent variable is students' reading proficiency, reflected by students' reading score in the PISA reading test.

of missing values was conducted in this study. Many researchers have advocated the use of missForest, a non-parametric method based on the randomForest model, in working with samples that involve different data types (Stekhoven and Bühlmann, 2012; Jin et al., 2015; Finch et al., 2016). Thus, because the sample included in this study contains both continuous and dichotomous variables, the missForest() function was used to impute the missing values.

Second, the correlation coefficients among the nine independent variables and ten PVs of reading performance were computed, and they were within the acceptable limits. Further, the T-value and the F-value needed to be emphasized to determine the correlation between nine independent variables and reading proficiency.

Third, the lm() function from the core package 'stats' was used to compute the MLR model. For each plausible value of reading performance, the model was built by the regressors and covariates.

The summary statistics of variables are presented in **Table 3**. As there were ten PVs, ten MLR models were eventually produced. The residuals (ε), estimates (β), intercept (β_0), standard error (SE), multiple R-squared (R^2) and p-values of the T-statistic and F-statistic are shown in the results for further discussion.

Fourth, assumptions of homoscedasticity and endogeneity were checked. Widely used to verify whether a regression model contains heteroskedastic error (Jeong and Lee, 2008), White's test (White, 1980) was applied in this study by the computing heteroscedasticity-robust standard error in test statistics (Wooldridge, 2003). Moreover, the problem of endogeneity might exist when the ICT use is an endogenous variable (Fariña et al., 2015). Therefore, the assumption of endogeneity was checked with all three covariates of ICT use, i.e., USESCH, HOMESCH and ENTUSE. The differences between two regression models of with and without any of these covariates for the rest of the variables were calculated and provided in **Supplementary Tables S1–S6**, respectively. The comparisons of result difference of each ICT use variable were presented in **Table 4**. No significant differences were found with and without these variables, respectively, in this process.

TABLE 3 | Descriptive statistics of ICT availability, ICT use, ICT attitudes and student background based on the PISA 2015 computer-based reading assessment.

Variable	Variable description	Mean	*SD*	Minimum	Maximum
ICT availability					
ICTHOME	ICT available at home index	8.600	1.638	0.000	11.000
ICTSCH	ICT available at school index	7.013	1.932	0.000	10.000
ICT use					
USESCH	Use of ICT at school in general	0.258	0.826	−1.668	3.629
HOMESCH	ICT use outside of school for schoolwork	−0.058	0.942	−2.691	3.604
ENTUSE	ICT use outside of school leisure	−0.010	0.888	−3.710	4.848
ICT attitudes					
INTICT	Students' ICT interest	0.131	0.935	−2.988	2.819
AUTICT	Students' perceived autonomy related to ICT use	0.130	0.900	−2.503	2.096
COMPICT	Students' perceived ICT competence	0.090	0.886	−2.706	2.074
SOIAICT	Students' ICT as a topic in social interaction	0.095	0.880	−2.136	2.428
Student background					
ESCS	Index of economic, social and cultural status	0.107	0.822	−4.692	3.567
Gender	Students' gender	/	/	0.000	1.000

Sources: OECD PISA 2015 general database. N = 37,155. The independent variables ICTHOME, ICTSCH, USESCH, HOMESCH, ENTUSE, INTICT, AUTICT, COMPICT, SOIAICT and the covariate ESCS were all derived variables based on IRT scaling. Gender is a binary variable, in which female is coded as "0" while male is coded as "1." The dependent variable is students' reading proficiency, reflected by students' reading score in the PISA reading test.

RESULTS

This article aimed to examine the influence of ICT impactors on students' reading proficiency in high-achieving countries; therefore, the five representative countries were assessed as a cohort with high-achieving reading proficiency.

Demographic Covariates

Regarding the PISA reading proficiency, the fundamental demographic factors involved the ESCS and gender (Petko et al., 2017; Hu et al., 2018). Thus, these two factors were included as the two demographic covariates in this study. Both ESCS (β = 47.930, *SE* = 0.663, $p < 0.001$) and gender (β = −28.506, *SE* = 1.039, $p < 0.010$) were significantly correlated with reading proficiency in **Table 5**. Specifically, ESCS was positively associated with the students' reading performance. For a one-point increment in ESCS, the students' reading scores increased by 39.398 points (β*SD), which demonstrated that the students in countries with higher ESCS tended to achieve better reading results.

Table 5 presents the results for all required coefficients for the statistically significantly related factors included in the optimal MLR model. As shown, the explained variance for the model varied from $R^2 = 0.209$ to $R^2 = 0.214$. In the fields of humanities and social sciences, these R^2 values were within an acceptable range because it was not expected that all relevant variables would be included to indicate the subjects' behavior. In the existing studies of regression analysis using the PISA dataset (e.g., Chiacchio et al., 2016; Naumann and Sälzer, 2017; Tay et al., 2017), the maximum R^2 reached 0.310, 0.239, and 0.230, respectively. Even if the R^2 was low in this study, the factors were significantly correlated, which means that important conclusions could still be drawn from the model (Neter et al., 2012). The detailed information of all statistical analyses conducted in this study are available upon request.

ICT-Related Factors

As shown in **Table 5**, ICT availability at home (β = −4.331, *SE* = 0.396, $p < 0.001$) and at school (β = −3.265, *SE* = 0.295, $p < 0.001$) was negatively associated with students' reading proficiency: with a one-point improvement in the availability of ICT at home and at school, students' reading scores decreased by −7.094 and −6.308 points (β*SD), respectively. Regarding use, ICT use at school in general (β = −7.536, *SE* = 0.779, $p < 0.001$) was negatively related to reading performance; reading scores were decreased by 6.225 points (β*SD), with one-point growth in the use of ICT at school. The use of ICT outside of school for entertainment (β = −8.148, *SE* = 0.746, $p < 0.001$) indicated a negative correlation with reading proficiency; the use of ICT outside of school for entertainment was increased by one point, and reading scores dropped by 7.236 points (β*SD). No significant association was found between the use of ICT outside of school for schoolwork and reading proficiency. With regard to students' attitudes toward ICT, all attitudinal factors examined were significantly related to reading performance: interest in ICT (β = 9.955, *SE* = 0.661, $p < 0.001$) and perceived autonomy related to ICT use (β = 23.529, *SE* = 0.775, $p < 0.001$) were positively related to reading scores, whereas perceived ICT competence (β = −2.931, *SE* = 0.796, $p < 0.001$) and enjoyment of social interactions through ICT (β = −16.001, *SE* = 0.709, $p < 0.001$) were negatively associated with reading performance. Specifically, reading scores increased by 9.308 and 21.076 (β*SD) points with every one-point increase in students' interest in ICT and perceived ICT autonomy, respectively. Conversely, with a one-point improvement in perceived ICT competence, students' reading score decreased by 2.597 points (β*SD). One point of growth in their enjoyment of ICT use for social interaction was found to reduce reading scores by 14.065 points (β*SD).

Moreover, the relationship between ICT impact factors and students' reading proficiency in each of the five performing

TABLE 4 | Comparison of the results of the regression models with and without each ICT use factor of USESCH, HOMESCH, and ENTUSE.

Factor		Regression model results with USESCH		Regression model results without USESCH		Differences		Regression model results with HOMESCH		Regression model results without HOMESCH		Differences		Regression model results with ENTUSE		Regression model results without ENTUSE		Differences
ICTHOME	β	−4.331*** (0.396)	β	−4.438*** (0.363)	β	0.107 (0.033)	β	−4.331*** (0.396)	β	−4.334*** (0.362)	β	0.003 (0.034)	β	−4.331*** (0.396)	β	−4.775*** (0.360)	β	0.444 (0.036)
	β*SD	−7.094	β*SD	−7.270	β*SD	0.176	β*SD	−7.094	β*SD	−7.050	β*SD	0.044	β*SD	−7.094	β*SD	−7.821	β*SD	0.727
ICTSCH	β	−3.265*** (0.295)	β	−3.829*** (0.288)	β	0.564 (0.007)	β	−3.265*** (0.295)	β	−3.268*** (0.294)	β	0.003 (0.001)	β	−3.265*** (0.295)	β	−3.131*** (0.296)	β	0.134 (0.001)
	β*SD	−6.308	β*SD	−7.214	β*SD	0.906	β*SD	−6.308	β*SD	−6.314	β*SD	0.006	β*SD	−6.308	β*SD	−6.049	β*SD	0.259
USESCH	β	/	β	/	β	/	β	−7.536*** (0.779)	β	−7.59*** (0.710)	β	0.014 (0.069)	β	−7.536*** (0.779)	β	−8.634*** (0.780)	β	1.098 (0.001)
	β*SD	/	β*SD	/	β*SD	/	β*SD	−6.225	β*SD	−5.286	β*SD	0.939	β*SD	−6.225	β*SD	−7.132	β*SD	0.907
HOMESCH	β	−0.325*** (0.700)	β	−3.127*** (0.640)	β	2.802 (0.060)	β	/	β	/	β	/	β	−0.325*** (0.700)	β	−8.179*** (0.728)	β	7.854 (0.028)
	β*SD	−0.306	β*SD	−1.245	β*SD	0.939	β*SD	/	β*SD	/	β*SD	/	β*SD	−7.235	β*SD	−7.263	β*SD	0.028
ENTUSE	β	−8.148*** (0.746)	β	−9.016*** (0.745)	β	0.868 (0.001)	β	−8.148*** (0.746)	β	−8.179*** (0.728)	β	0.031 (0.042)	β	/	β	/	β	/
	β*SD	−7.236	β*SD	-8.002	β*SD	0.766	β*SD	−7.236	β*SD	−7.263	β*SD	0.027	β*SD	/	β*SD	/	β*SD	/
INTICT	β	9.955*** (0.661)	β	9.827*** (0.662)	β	0.128 (0.001)	β	9.955*** (0.661)	β	9.954*** (0.661)	β	0.010 (0.00)	β	9.955*** (0.661)	β	8.513*** (0.648)	β	1.442 (0.013)
	β*SD	9.308	β*SD	9.190	β*SD	0.118	β*SD	9.308	β*SD	9.307	β*SD	0.001	β*SD	9.308	β*SD	8.359	β*SD	0.949
AUTICT	β	23.529*** (0.775)	β	23.673*** (0.776)	β	0.144 (0.001)	β	23.529*** (0.775)	β	23.533*** (0.774)	β	0.004 (0.001)	β	23.529*** (0.775)	β	23.630*** (0.772)	β	0.101 (0.003)
	β*SD	21.076	β*SD	21.464	β*SD	0.388	β*SD	21.076	β*SD	21.180	β*SD	0.104	β*SD	21.076	β*SD	20.367	β*SD	0.709
COMPICT	β	−2.931*** (0.796)	β	−3.206*** (0.794)	β	0.275 (0.002)	β	−2.931*** (0.796)	β	−2.934*** (0.794)	β	0.003 (0.002)	β	−2.931*** (0.796)	β	−3.208*** (0.787)	β	0.277 (0.009)
	β*SD	−2.597	β*SD	−2.844	β*SD	0.247	β*SD	−2.597	β*SD	−2.600	β*SD	0.003	β*SD	−2.497	β*SD	−1.559	β*SD	0.938
SOIAICT	β	−16.001*** (0.709)	β	−16.321*** (0.710)	β	0.320 (0.001)	β	−16.001*** (0.709)	β	−16.014*** (0.705)	β	0.013 (0.004)	β	−16.001*** (0.709)	β	−16.937*** (0.705)	β	0.936 (0.004)
	β*SD	−14.065	β*SD	−14.351	β*SD	0.286	β*SD	−14.065	β*SD	−14.076	β*SD	0.011	β*SD	−14.065	β*SD	−14.887	β*SD	0.822
ESCS	β	47.930*** (0.663)	β	−47.644*** (0.665)	β	0.286 (0.002)	β	47.930*** (0.663)	β	47.920*** (0.660)	β	0.010 (0.003)	β	47.930*** (0.663)	β	48.320*** (0.664)	β	0.390 (0.001)
	β*SD	39.398	β*SD	39.146	β*SD	0.252	β*SD	39.398	β*SD	39.396	β*SD	0.002	β*SD	39.398	β*SD	39.719	β*SD	0.321
Gender (female = 0)	β	−28.506*** (1.039)	β	−28.363*** (1.040)	β	0.143 (0.001)	β	−28.506*** (1.039)	β	−28.488*** (1.035)	β	0.018 (0.004)	β	−28.506*** (1.039)	β	−29.966*** (0.826)	β	1.460 (0.213)
	β*SD	−14.253	β*SD	−14.181	β*SD	0.072	β*SD	−14.253	β*SD	−14.244	β*SD	0.009	β*SD	−14.253	β*SD	−14.983	β*SD	0.730

The coefficient of the regression model presented in this table were the mean coefficient of the 10 models. Heteroscedasticity-robust standard errors are listed in parentheses. Results of the model with and without USESCH, HOMESCH and ENTUSE were compared in ***Supplementary Tables S2****,* ***S4****,* ***S6****, respectively, with the main indicator of β*SD. Significant codes: ***$p < 0.05$.*

TABLE 5 | The effect of ICT impact factors on reading proficiency.

Factor	Model 1 PV1READ		Model 2 PV2READ		Model 3 PV3READ		Model 4 PV4READ		Model 5 PV5READ		Model 6 PV6READ		Model 7 PV7READ		Model 8 PV8READ		Model 9 PV9READ		Model 10 PV10READ		Mean	
	β	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**	**β**	**β*SD**
ICT availability																						
ICTHOME	−4.434*** (0.700)	−7.263	−4.162*** (0.362)	0.333	−4.721*** (0.358)	0.331	−3.937*** (0.361)	0.332	−4.209*** (0.368)	0.335	−3.869*** (0.361)	0.330	−4.383*** (0.364)	0.333	−4.306*** (0.360)	0.332	−4.456*** (0.363)	0.334	−4.832*** (0.362)	0.334	−4.331*** (0.396)	−7.094
ICTSCH	−3.243*** (0.293)	−6.265	−3.313*** (0.295)	0.271	−2.972*** (0.292)	0.270	−3.131*** (0.296)	0.271	−3.047*** (0.296)	0.274	−3.493*** (0.291)	0.270	−3.535*** (0.297)	0.232	−3.684*** (0.294)	0.271	−3.030*** (0.295)	0.273	−3.201*** (0.298)	0.273	−3.265*** (0.295)	−6.308
ICT use																						
USESCH	−7.20*** (0.782)	−5.947	−7.987*** (0.781)	0.674	7.275*** (0.764)	0.671	−8.042*** (0.783)	0.673	−6.425*** (0.784)	0.680	−7.673*** (0.774)	0.670	−7.394*** (0.780)	0.744	−7.807*** (0.782)	0.742	−7.489*** (0.776)	0.747	−8.070*** (0.787)	0.677	−7.536*** (0.779)	−6.225
HOMESCH	−0.246 (0.700)	−0.232	−0.464 (0.702)	0.660	−0.574 (0.964)	0.657	−0.193 (0.695)	0.658	0.300 (0.705)	0.665	−0.144 (0.695)	0.655	−0.989 (0.696)	0.660	0.994 (0.705)	0.657	−1.272 (0.702)	0.662	−0.663 (0.708)	0.663	−0.325 (0.700)	−0.306
ENTUSE	−8.884*** (0.746)	−7.89	−8.102*** (0.752)	0.662	−8.190*** (0.741)	0.660	−7.654*** (0.736)	0.661	−8.083*** (0.7540	0.667	−7.973*** (0.745)	0.658	−7.768*** (0.740)	0.680	−8.285*** (0.759)	0.678	−8.002*** (0.739)	0.683	−8.541*** (0.752)	0.665	−8.148*** (0.746)	−7.236
ICT attitudes																						
INTICT	9.821*** (0.662)	9.183	9.376*** (0.666)	0.637	10.249*** (0.658)	0.635	9.630*** (0.657)	0.636	9.787*** (0.663)	0.643	10.084*** (0.654)	0.633	9.821*** (0.665)	0.637	10.519*** (0.660)	0.636	10.017*** (0.664)	0.630	10.242*** (0.658)	0.640	9.955*** (0.661)	9.308
AUTICT	23.055*** (0.767)	20.795	24.482*** (0.778)	0.750	22.940*** (0.770)	0.747	23.826*** (0.771)	0.748	23.504*** (0.782)	0.756	22.997*** (0.770)	0.745	23.286*** (0.778)	0.750	23.903*** (0.776)	0.748	23.221*** (0.773)	0.753	24.075*** (0.780)	0.753	23.529*** (0.775)	21.076
COMPICT	−2.874*** (0.790)	−2.546	−3.015*** (0.804)	0.766	−2.238*** (0.790)	0.763	−3.243*** (0.7880	0.764	−3.043*** (0.799)	0.772	−2.404*** (0.792)–	0.761	−1.831*** (0.800)	0.765	−3.986*** (0.795)	0.763	−3.105*** (0.794)	0.768	−3.572*** (0.792)	0.769	−2.931*** (0.796)	−2.597
SOIAICT	−15.693*** (0.709)	−13.794	−16.337*** (0.711)	0.687	−16.752*** (0.703)	0.685	−16.192*** (0.705)	0.684	−16.055*** (0.717)	0.693	−15.743*** (0.700)	0.682	−15.895*** (0.711)	0.690	−16.064*** (0.713)	0.688	−15.525*** (0.711)	0.693	−15.755*** (0.713)	0.690	−16.001*** (0.709)	−14.065
Student background																						
ESCS	48.210*** (0.656)	39.629	47.380*** (0.667)	0.631	47.676*** (0.660)	0.628	48.116*** (0.664)	0.630	47.640*** (0.673)	0.636	47.767*** (0.656)	0.626	47.884*** (0.666)	0.632	47.885*** (0.663)	0.630	48.188*** (0.661)	0.635	48.549*** (0.665)	0.633	47.930*** (0.663)	39.398
Gender (female = 0)	−28.410*** (1.035)	−14.205	−27.647*** (1.040)	1.020	−27.441*** (1.037)	1.017	−29.484*** (1.037)	1.018	−28.844*** (1.047)	1.029	−28.933*** (1.032)	1.013	−30.656*** (1.039)	1.023	−27.921*** (1.035)	1.021	−28.661*** (1.043)	1.028	−27.062*** (1.044)	1.025	−28.506*** (1.039)	−14.253

*N = 37,155. The dependent variable is students' readings score. Models 1 to 10 refer to the regression models for ten plausible values of reading score. In the PISA 2015, each student has 10 plausible values of reading scores (PV1READ∼PV10READ). A higher plausible value reflects a higher reading proficiency. The regression model is estimated using Equation. Since the independent variables were derived based on IRT scaling, with one percent change in the independent variable, the dependent variable is changed by the coefficient multiplied by its standard deviation (β*SD). Heteroscedasticity-robust standard errors are listed in parentheses. Significant codes: ***$p < 0.001$.*

countries was also investigated through the same procedure, respectively, in **Table 6**. As shown, the ICT availability at home (ICTHOME) and the gender (Gender) remained negatively associated with students' reading proficiency in each of the five counties, and the interest in ICT (INTICT) and the ESCS remained positively correlated with students' reading proficiency in each of the five counties. For the remaining factors, they were differently associated with reading proficiency among different countries. These results indicated that: four factors (i.e., ICTHOME, INTICT, ESCS, and Gender) were simultaneously identified for all five countries as closely relevant to the students' reading proficiency, whereas the other factors were differently associated with students' reading proficiency among countries. For example, the ICT availability at school (ICTSCH) was negatively associated with reading proficiency in Australia ($p = 0.021$, $\beta = -1.296$, $SE = 0.562$), France ($p < 0.001$, $\beta = -9.138$, $SE = 0.725$) and New Zealand ($p < 0.001$, $\beta = -3.478$, $SE = 0.978$). The correlation was insignificant in Finland (p = 0.098) and Singapore (p = 0.068).

DISCUSSION

The Availability of ICT

The availability of ICT includes ICT availability at home (ICTHOME) and ICT availability at school (ICTSCH) (OECD, 2016). On one hand, ICTHOME is found to be inversely related to students' reading achievement in high-achieving countries, which is consistent with the previous research (Lee and Wu, 2013). This finding might be explained by the low quality of students' ICT use at home without proper guidance and timely supervision from their parents. Students with access to ICT devices at home (e.g., computers, cell phones, e-books, printers, portable music players) do possess more computer skills (Kuhlemeier and Hemker, 2007) and tend to perform better on reading when assessed by computer (Rasmusson and Åberg-Bengtsson, 2015). However, the overuse or abuse of ICT tends to form detrimental habits such as addiction to computer games, which in turn lowers reading proficiency (Rasmusson and Åberg-Bengtsson, 2015). Hence, parents are suggested to carefully monitor their children's access to ICT facilities at home and to appropriately direct them to utilize online resources in a reasonable way (Lee and Wu, 2012). On the other hand, ICTSCH is negatively correlated with students' reading performance in this study, which is consistent with Lai's (2016) study. This result is closely related to ICT use at school, which is discussed in detail in the next section.

The Use of ICT

The use of ICT contains ICT use at school in general (USESCH), ICT use at home for schoolwork (HOMESCH), and ICT use at home for leisure (ENTUSE) (OECD, 2016). ICT use at school is negatively related to students' reading scores, which is consistent with the findings of previous studies (Petko et al., 2017; Tay et al., 2017; Hu et al., 2018). During the process of using ICT in everyday education, teachers may encounter a number of barriers. Ertmer (1999) classified these barriers into two categories: extrinsic and intrinsic barriers. Extrinsic barriers include lack of access, time, support, resources and training, and intrinsic barriers include attitudes, beliefs, practices and resistance. In terms of intrinsic barriers with regard to teachers' preparedness and perception, although teachers believe ICT use in education is beneficial and may be able to adeptly use the Internet, e-mail, Microsoft Word and PowerPoint for reading teaching, they might possess only limited knowledge in using ICT for more advanced functions, e.g., spreadsheets, concept mapping, programing languages, multimedia authoring and modeling software to compose adapted teaching materials or tailored approaches for students with different reading levels. This indicates a situation where the use of ICT in class is restricted to basic pedagogical practices rather than being effectively integrated into the school curriculum (Aydin, 2013). Therefore, schools are supposed to organize training programs to equip teachers with important ICT knowledge and sufficient ICT skills as well as provide in-time technical support once teachers encounter any difficulty in using ICT in class and so forth (Hadi and Zeinab, 2012). In this case, teachers would be able to use ICT as cognitive tools in class, contributing to an ideal technology-assisted learning environment (e.g., Kommers et al., 2001; Nissen and Tea, 2012; Wei and Hu, 2018; Wei et al., 2018).

The results regarding the influence of the use of ICT for academic purposes outside of school on reading proficiency have varied across the previous studies. In this study, no significant connection is found between ICT use at home for schoolwork and reading proficiency. In the existing studies, Petko et al. (2017) discovered that ICT use for schoolwork outside of school is positively associated with students' reading performance, which aligns with the research finding of Skryabin et al. (2015). In contrast, Gumus and Atalmis (2011) discovered the negative relationship of ICT academic use at home. These conflicting results might be explained by the fact that the PISA ICT questionnaire have changed over time, as explained in the introduction. In detail, Skryabin et al. (2015) and Petko et al. (2017) applied the ICT questionnaire in the PISA 2012, at which time the index for ICT use outside of school for academic purposes was determined by seven measurements. In Gumus and Atalmis (2011) study, this index was based on five questions in the PISA 2006 ICT questionnaire (OECD, 2006). However, in the current study, the final index of ICT use outside of school for schoolwork is derived from twelve indexes in the PISA 2015 ICT questionnaire, including all seven indexes that were examined in the PISA 2012 (OECD, 2017).

In this study, ICT use outside of school for entertainment is found to be inversely correlated with reading proficiency, which contradicts the findings of some of the past studies. For instance, Gumus and Atalmis (2011) proposed that using ICT devices for leisure, such as playing computer games, may alleviate Turkish students' stress, increase their momentum, and inspire them to learn more efficiently. However, the pattern of a negative correlation between ICT use outside of school for entertainment and reading performance is found in high-achieving countries (Woessmann and Fuchs, 2005; OECD, 2006, 2015; Petko et al., 2017). Another possible explanation might be the opportunity cost of spending most of the time online outside school for entertainment rather

TABLE 6 | The effect of ICT impact factors on reading proficiency in each of the five countries.

Group of factors	Measurement	Australia	Finland	France	New Zealand	Singapore	All five countries
ICT availability							
ICTHOME	*p*-value	0.000	0.000	0.002	0.001	0.000	0.000
	β	−2.307	−6.68	−3.163	−3.425	−3.353	−4.331
	Robust SE	0.611	0.859	1.001	1.053	0.717	0.396
	β*SD	−3.576	−10.354	−4.903	−5.309	−5.197	−7.094
ICTSCH	*p*-value	0.021	**0.098**	0.000	0.000	**0.068**	0.000
	β	−1.296	1.008	−9.138	−3.478	−1.091	−3.265
	Robust SE	0.562	0.609	0.725	0.978	0.598	0.295
	β*SD	0.034	1.635	−14.822	−5.641	−1.770	−6.308
ICT use							
USESCH	*p*-value	0.000	0.000	**0.275**	0.000	0.000	0.000
	β	5.818	−15.801	1.768	−16.562	−8.532	−7.536
	Robust SE	1.577	1.938	1.619	2.685	1.561	0.779
	β*SD	4.311	−11.709	1.310	−12.272	−6.322	−6.225
HOMESCH	*p*-value	**0.866**	0.010	0.001	0.008	0.000	0.468
	β	0.213	−3.870	−5.651	6.207	6.432	−0.325
	Robust SE	1.262	1.505	1.647	2.337	1.719	0.700
	β*SD	0.193	−3.510	−5.125	5.630	5.834	−0.306
ENTUSE	*p*-value	0.000	**0.226**	0.000	0.000	**0.072**	0.000
	β	−18.430	−2.003	−5.965	−8.780	−3.418	−8.148
	Robust SE	1.289	1.655	1.657	2.294	1.901	0.746
	β*SD	−16.163	−1.757	−5.231	−7.700	−2.998	−7.236
ICT attitudes							
INTICT	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	β	14.716	7.875	0.173	12.993	8.274	9.955
	Robust SE	1.182	1.617	1.485	1.967	1.425	0.661
	β*SD	13.038	6.977	0.153	11.512	7.331	9.308
AUTICT	*p*-value	0.000	0.000	**0.907**	0.000	0.000	0.000
	β	22.733	15.973	20.775	18.673	22.32	23.529
	Robust SE	1.350	1.662	1.941	2.194	1.573	0.775
	β*SD	20.028	14.072	18.303	16.451	19.664	21.076
COMPICT	*p*-value	**0.616**	**0.461**	**0.195**	**0.202**	0.001	0.003
	β	0.703	−1.311	2.251	2.967	−6.138	−2.931
	Robust SE	1.403	1.779	1.735	2.326	1.802	0.796
	β*SD	0.595	−1.109	1.904	2.510	−5.193	−2.597
SOCIAICT	*p*-value	0.000	**0.965**	0.000	0.000	0.000	0.000
	β	−22.012	−0.070	−14.048	−23.675	−21.634	−16.001
	Robust SE	1.270	1.599	1.556	2.208	1.526	0.709
	β*SD	−18.204	−0.058	−11.618	−19.579	−17.891	−14.065
Student background							
ESCS	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	β	43.679	40.262	57.960	49.631	45.237	48.549
	Robust SE	1.112	1.580	1.667	1.990	1.396	0.665
	β*SD	35.293	32.532	46.832	40.102	36.551	39.398
Gender	*p*-value	0.000	0.000	0.000	0.000	0.000	0.000
	β	−26.038	−46.999	−22.714	−31.721	−14.157	−28.506
	Robust SE	1.696	2.348	2.589	3.032	2.338	1.039
	β*SD	−13.019	−23.500	−11.357	−15.861	−7.079	−14.253

Coefficients that resulted significant considering a 0.05 significance level appear in bold. Heteroscedasticity-robust standard error was computed to test the potential heteroscedasticity.

than spending that time reading (Petko et al., 2017). Educators should devote more effort to monitoring and evaluating students' reading strategies to achieve meaningful e-teaching outcomes. However, ICT use at school in Australia is positively correlated with students' reading proficiency, which might be caused by the education policies in Australia. These policies contribute a lot to the effectively use of ICT in schools (Radhika and Wu, 2015).

Attitudes Toward ICT

Regarding attitudes toward ICT, two attitudinal factors, i.e., students' interest and perceived autonomy in using ICT, are closely associated with the reading proficiency in high-performing countries in this study. This finding is novel, as few studies have confirmed the predominant significant role of ICT-related motivation and self-efficacy in reading scores beyond students' capabilities. The previous studies have found that the impact of these two attitudinal factors on students' achievement scores is complex (e.g., Papanastasiou et al., 2004; Lee and Wu, 2012). Lee and Wu (2012) observed that students' perceptions of educational technology were positively correlated with their academic performance based on the PISA 2009 dataset whereas Papanastasiou et al. (2004) suggested a negative correlation. The reason for the fundamental influence of interest might be the digital learning potential reflected by the items measuring students' interest in ICT in the PISA 2015 ICT familiarity questionnaire. This potential is measured by two main items: (1) The Internet is a great resource for obtaining information in which I am interested, and (2) I am really excited about discovering new digital devices or applications (OECD, 2017). In effect, these two questions reflect students' acceptance of ICT related technology. ICT has brought tremendous change by offering readers the opportunity to engage in more flexible reading activities via computers. Nonetheless, many adolescents born in the 1990s have uninterested, skeptical or even fearful attitudes toward e-learning because of the complicated and misleading navigation, non-intuitive design, and user-unfriendly operations, which might hinder their access to informative resources (Hyman et al., 2014). This attitude of rejection decreases students' autonomy in utilizing ICT facilities for learning. Students without an interest in applying ICT to help them with their work are unlikely to delve into the manuals of electronic devices, choose helpful applications or install updated learning software independently. Hence, students' indifference to ICT shows little possibility for automatic e-learning in further study, which may hinder their reading performance. This interpretation seems plausible in light of the previous studies on the gender gap in online reading, which observe that the advantage in female students over their male counterparts in paper-based reading decreases when they read online. Based on Bandura's self-efficacy theory (1993), it is possible that boys' greater interest and girls' higher anxiety in the electronic reading environment contributed to the smaller gender gap in digital reading (Nele and Franziska, 2019). Therefore, new effective and technological tools used in the classroom should be geared to students' interest; in particular, attractive educational applications could trigger students' positive attitudes or behavior in class (Mera et al., 2019).

With regard to students' perceived ICT competence in using digital devices, this study finds a slightly negative association. In the meantime, students' ICT use for social interaction is negatively correlated with reading proficiency in the sample countries. In the PISA 2015 ICT familiarity questionnaire, the questions on this index can be generalized into two categories. One category is ICT as a theme of social communication, and the other is ICT use for social interaction. Although students might receive assistance in using digital devices from social media, using ICT for social communication exerts a greater negative correlation with reading proficiency. This result is consistent with those of previous studies (e.g., Fox et al., 2009; Jacobsen and Forste, 2011) that confirmed that concurrent ICT use for social communication and for reading were negatively associated with their efficiency. Jacobsen and Forste (2011) further proposed that the metacognitive mechanism behind this negative correlation is the distraction of attention and the impairment of short-term memory when performing multiple tasks. Additionally, this finding further explains the negative impact of ICT use at school as mentioned above. Areepattamannil and Khine (2017) revealed the close connection between the frequency of ICT use for social interaction and ICT use at school. In this case, the fact that ICT's use at school is negatively associated with reading scores is attributed not only to teachers' behavior but also to students' reading activities at school. To solve this problem, appropriate direction and timely scrutiny are necessary to prevent students from becoming obsessed with online entertainment such as playing computer games and engaging in social networking activities. The significant negative impact of social interaction activities on students' reading proficiency in high-achieving countries reflects the fact that social media addiction poses a great threat to reading proficiency with the popularity of ICT.

CONCLUSION

This study used multiple linear regression models to analyze the relationship between ICT impact factors and early adolescents' reading proficiency in five countries with extremely high reading proficiency. It was found that students' attitudes toward ICT including interest levels and perceived autonomy contributed most to students' high reading proficiency, rather than ICT availability or ICT use. The current study makes the following three primary contributions to the field: (a) This study delves into the association between the proposed ICT-related factors and students' reading proficiency in the context of representative countries with excellent reading proficiency based on the latest PISA dataset, and it makes reasonable inferences for illustration; (b) The study reflects upon the application of educational technology in an ICT-assisted learning environment and gives constructive advice with regard to the findings; and (c) Based on the previous literature, this study offers a comprehensive overview of how ICT influences reading performance.

Future research should address a few suggestions. First, considering the exploratory nature of the research, if possible, in the future, longitudinal research can expand the scale of the research. Second, since most of the questions in the PISA questionnaire were the self-reported answers of students, the endogeneity of variables might be a problem. In a pioneering PISA study (Fariña et al., 2015), the Hausman test (Hausman, 1978) was used to diagnose the appropriateness of the endogeneity assumption. Furthermore, propensity score matching approach can be applied to avoid the self-selection problem and obtain an unbiased sample (e.g., Crespo-Cebada et al., 2014). Although this problem does not exist in this study, it still deserves special attention in future research. Additionally, the application of more advanced statistical model, for instance, a linear mixed-effects model (e.g., Hesselmann, 2018), is also essential for future PISA-based studies.

DATA AVAILABILITY

The data that support the findings of this study are available at http://www.oecd.org/pisa/data/. This is public data released by the OECD.

ETHICS STATEMENT

This study was approved by the Research Ethics Board of Zhejiang University and granting agency, and was performed in accordance with the relevant guidelines and regulations.

AUTHOR CONTRIBUTIONS

YX designed the study, analyzed and interpreted the data, and wrote and revised the manuscript. YL revised the manuscript. JH supervised the study, designed the study, interpreted the data, and wrote and revised the manuscript.

FUNDING

This study was partially funded by the National Education Science Youth Project for the 13th Five-Year Plan (Grant No. CIA170274).

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their constructive comments on this manuscript. The authors also express sincere gratitude to the OECD for its generous publication of the data and the valuable suggestions on the methods of addressing the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01646/full#supplementary-material

REFERENCES

Areepattamannil, S., and Khine, M. S. (2017). Early adolescents' use of information and communication technologies (ICTs) for social communication in 20 countries: examining the roles of ICT-related behavioral and motivational characteristics. *Comput. Hum. Behav.* 73, 263–272. doi: 10.1016/j.chb.2017.03.058

Aydin, S. (2013). Teachers' perceptions about the use of computers in EFL teaching and learning: the case of Turkey. *Comput. Assist. Lang. Learn.* 26, 214–233. doi: 10.1080/09588221.2012.654495

Chen, J., and Hu, J. (2018). Enhancing L2 learners' critical thinking skills through a connectivism-based intelligent learning system. *Int. J. Engl. Linguist.* 8, 12–21. doi: 10.5539/ijel.v8n6p12

Chen, J., Zhang, Y., Wei, Y., and Hu, J. (2019). Discrimination of the contextual features of top performers in scientific literacy using a machine learning approach. *Res. Sci. Educ.* [Preprint]. doi: 10.1007/s11165-019-9835-y

Cheung, K. C., Mak, S. K., and Sit, P. S. (2013). Online reading activities and ICT use as mediating variables in explaining the gender difference in digital reading literacy: comparing hong kong and Korea. *Asia-Pac. Educ. Res.* 22, 709–720. doi: 10.1007/s40299-013-0077-x

Chiacchio, C. D., Stasio, S. D., and Fiorilli, C. (2016). Examining how motivation toward science contributes to omitting behaviors in the Italian PISA 2006 sample. *Learn. Individ. Differ.* 50, 56–63. doi: 10.1016/j.lindif.2016.06.025

Coiro, J., and Dobler, E. (2007). Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the internet. *Read. Res. Q.* 42, 214–257. doi: 10.1598/RRQ.42.2.2

Crespo-Cebada, E., Pedraja-Chaparro, F., and Santin, D. (2014). Does school ownership matter? An unbiased efficiency comparison for regions of Spain. *J. Prod. Anal.* 41, 153–172. doi: 10.1007/s11123-013-0338-y

Cui, X. J., Yang, Q. X., Li, B., Tang, J., Zhang, X. Y., Li, S., et al. (2019). Assessing the effectiveness of direct data merging strategy in long-term and large-scale pharmacometabonomics. *Front. Pharmacol.* 10:127. doi: 10.3389/fphar.2019.00127

Deng, Z., and Gopinathan, S. (2016). PISA and high-performing education systems: explaining Singapore's education success. *Comput. Educ.* 52, 449–472. doi: 10.1080/03050068.2016.1219535

Eklund, K., Torppa, M., Sulkunen, S., Niemi, P., and Ahonen, T. (2018). Early cognitive predictors of PISA reading in children with and without family risk for dyslexia. *Learn. Individ. Differ.* 64, 94–103. doi: 10.1016/j.lindif.2018.04.012

Erdogdu, F., and Erdogdu, E. (2015). The impact of access to ICT, student background and school/home environment on academic success of students in Turkey: an international comparative analysis. *Comput. Educ.* 82, 26–49. doi: 10.1016/j.compedu.2014.10.023

Ertmer, P. (1999). Addressing first- and second -order barriers to change: strategies for technology intergration. *Educ. Technol. Res. Dev.* 47, 47–61. doi: 10.1007/BF02299597

Fariña, P., San Martín, E., Preiss, D. D., Claro, M., and Jara, I. (2015). Measuring the relation between computer use and reading literacy in the presence of endogeneity. *Comput. Educ.* 80, 176–186. doi: 10.1016/j.compedu.2014.08.010

Finch, W. H., Finch, M. E. H., and Singh, M. (2016). Data imputation algorithms for mixed variable types in large scale educational assessment: a comparison of random forest, multivariate imputation using chained equations, and MICE with recursive partitioning. *Int. J. Quant. Res. Educ.* 3, 129–153. doi: 10.1504/IJQRE.2016.077803

Fox, A. B., Rosen, J., and Crawford, M. (2009). Distractions: does instant messaging affect college students' performance on a concurrent reading comprehension task? *Cyberpsychol. Behav.* 12, 51–53. doi: 10.1089/cpb.2008.0107

Gan, B., Menkhoff, T., and Smith, R. (2015). Enhancing students' learning process through interactive digital media: new opportunities for collaborative learning. *Comput. Hum. Behav.* 51, 652–663. doi: 10.1016/j.chb.2014.12.048

Gotoh, H., Murota, M., and Kamiyama, A. (2013). A cross-national analysis of parental involvement and student literacy. *Int. J. Comp. Sociol.* 4, 246–266. doi: 10.1177/0020715213501183

Gumus, S., and Atalmis, E. H. (2011). Exploring the relationship between purpose of computer usage and reading skills of Turkish students: evidence from PISA 2006. *Turk. Online J. Educ.* 10, 129–140. doi: 10.1080/1475939X.2011.588414

Hadi, S., and Zeinab, S. (2012). Challenges for using ICT in education: teachers' insights. *Int. J. e-Educ. e-Busi. e-Mana. e-Learn.* 2, 40–43.

Hausman, J. (1978). Specification test in econometrics. *Econometrica* 46, 1251–1271. doi: 10.2307/1913827

Hemmerechts, K., Agirdag, O., and Kavadias, D. (2016). The relationship between parental literacy involvement, socio-economic status and reading literacy. *Educ. Rev.* 69, 85–101. doi: 10.1080/00131911.2016.1164667

Hesselmann, G. (2018). Applying linear mixed effects models (LMMs) in within-participant designs with subjective trial-based assessments of awareness-a caveat. *Front. Psychol.* 9:788. doi: 10.3389/fpsyg.2018.00788

Holmes, W., and Rinaman, W. (2015). *Multiple Linear Regression. Statistical Literacy for Clinical Practitioners*. New York, NY: Springer International Publishing, 367–396.

Hu, J., and Wei, Y. (2018). Review of creativity and english language teaching: from inspiration to implementation. *Engl. Today* 35, 60–62. doi: 10.1017/S0266078418000299

Hu, X., Gong, Y., Lai, C., and Leung, F. K. S. (2018). The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: a multilevel analysis. *Comput. Educ.* 125, 1–13. doi: 10.1016/j.compedu.2018.05.021

Hyman, J. A., Moser, M. T., and Segala, L. N. (2014). Electronic reading and digital library technologies: understanding learner expectation and usage intent for mobile learning. *Educ. Technol. Res. Dev.* 62, 35–52. doi: 10.1007/s11423-013-9330-5

Jacobsen, W. C., and Forste, R. (2011). The wired generation: academic and social outcomes of electronic media use among university students. *Cyberpsychol. Beh. Soc. N.* 14, 275–280. doi: 10.1089/cyber.2010.0135

Jeong, J., and Lee, K. (2008). Bootstrapped White's test for heteroskedasticity in regression models. *Econ. Lett.* 63, 261–267. doi: 10.1016/S0165-1765(99)00036-1

Jerrim, J. (2015). Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent. *Oxford Rev. Educ.* 41, 310–333. doi: 10.1080/03054985.2015.1028525

Jin, Y., Li, B., Chen, N., Li, X., and Hu, J. (2015). The discrimination of learning styles by bayes-based statistics: an extended study on ILS system. *Control Intel. Syst.* 43, 68–75. doi: 10.2316/Journal.201.2015.2.201-2666

Kommers, P. A. M., Jonassen, D. H., and Mayes, J. T. (2001). *Cognitive Tools for Learning*. New York, NY: Springer-Verlag.

Kuhlemeier, H., and Hemker, B. (2007). The impact of computer use at home on students' internet skills. *Comput. Educ.* 49, 460–480. doi: 10.1016/j.compedu.2005.10.004

Lai, Y. H. (2016). Investigation on the relationship between information communication technology and reading literacy for northeast asian students. *MATEC Web Conf.* 71:03007. doi: 10.1051/matecconf/20167103007

Lee, Y. H., and Wu, J. Y. (2012). The effect of individual differences in the inner and outer states of ICT on engagement in online reading activities and PISA 2009 reading literacy: exploring the relationship between the old and new reading literacy. *Learn. Individ. Differ.* 22, 336–342. doi: 10.1016/j.lindif.2012.01.007

Lee, Y. H., and Wu, J. Y. (2013). The indirect effects of online social entertainment and information seeking activities on reading literacy. *Comput. Educ.* 67, 168–177. doi: 10.1016/j.compedu.2013.03.001

Li, J., and Chun, C. K. (2012). Effects of learning strategies on student reading literacy performance. *Read. Matrix* 12, 30–38.

Liberto, A. D. (2014). Length of stay in the host country and educational achievement of immigrant students. *Int. J. Manpow.* 36, 585–618. doi: 10.1108/IJM-11-2013-0261

Liu, Z. M. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. *J. Doc.* 61, 700–712. doi: 10.1108/00220410510632040

Mantoro, T., Fitri, E. M., Rusdah, R., Ayu, M. A., and Usino, W. (2017). The impact of information and communication technology (ICT) toward learning process and students' attitudes. *Adv. Sci. Lett.* 23, 844–847. doi: 10.1166/asl.2017.7554

Meng, L., Muñoz, M., Hess, K. K., and Liu, S. (2017). Effective teaching factors and student reading strategies as predictors of student achievement in PISA 2009: the case of China and the United States. *Educ. Rev.* 69, 1–17. doi: 10.1080/00131911.2016.1155537

Mera, C., Ruiz, G., Aguilar, M., Aragón, E., Delgado, C., Menacho, I., et al. (2019). Coming together: R&D and children's entertainment company in designing apps for learning early Math. *Front. Psychol.* 9:2751. doi: 10.3389/fpsyg.2018.02751

Mok, S. Y., Martiny, S. E., Gleibs, I. H., Keller, M. M., and Froehlich, L. (2016). The relationship between ethnic classroom composition and Turkish-origin and German students' reading performance and sense of belonging. *Front. Psychol.* 7:1071. doi: 10.3389/fpsyg.2016.01071

Montgomery, D. C., and Runger, G. C. (2007). *Applied Statistics and Probability for Engineers*. New York, NY: John Wiley and Sons, Inc.

Naumann, J. (2015). A model of online reading engagement: linking engagement, navigation, and performance in digital reading. *Comput. Hum. Behav.* 53, 263–277. doi: 10.1016/j.chb.2015.06.051

Naumann, J., and Sälzer, C. (2017). Digital reading proficiency in German 15-year olds: evidence from PISA 2012. *Z. Erziehwiss.* 20, 585–603. doi: 10.1007/s11618-017-0758-y

Nele, M., and Franziska, S. (2019). Gender gap in reading digitally? Examining the role of motivation and self-concept. *J. Educ. Res. Online* 11, 145–165. doi: 10.1016/j.chb.2015.06.051

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (2012). *Applied Linear Statistical Model*. Washington, DC: American Statistical Association.

Nissen, E., and Tea, E. (2012). Going blended: new challenges for second generation L2 tutors. *Comput. Assist. Lang. Learn.* 25, 145–163. doi: 10.1080/09588221.2011.636052

OECD (2006). *Are Students Ready for a Technology-Rich World?: What PISA Studies Tell Us*. Paris: OECD Publishing.

OECD (2009). *PISA Data Analysis Manual*, 2nd Edn. Paris: OECD Publishing.

OECD (2011). *PISA 2009 Results: Students Online: Digital Technologies and Performance*, Vol. VI. Paris: OECD Publications.

OECD (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

OECD (2015). *Students, Computers and Learning: Making the Connection*. Paris: OECD Publishing.

OECD (2016). *PISA 2015 Results Excellence and Equity in Education*, Vol. I. Paris: OECD Publishing.

OECD (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

Papanastasiou, E. C., Zembylas, M., and Vrasidas, C. (2004). "Reexamining patterns of negative computer-use and achievement relationships. Where and why do they exist?," in *Proceedings of the IRC-2014, TIMSS*, ed. C. Papanastasiou (Nicosia, CY: IEA-ETS Research Institute), 127–138.

Petko, D., Cantieni, A., and Prasse, D. (2017). Perceived quality of educational technology matters: a secondary analysis of students' ICT use, ICT-related attitudes, and PISA 2012 test scores. *J. Educ. Comput. Res.* 54, 1070–1091. doi: 10.1177/0735633116649373

Puteh, M., Zin, Z. M., and Ismail, I. (2016). Reading performance of Malaysian students across gender in PISA 2012. 3L-Lang. *Linguist. Lit.* 22, 109–121. doi: 10.17576/3L-2016-2202-08

Radhika, G., and Wu, M. (2015). Leaning too far? PISA, policy and Australia's top "five" ambitions. *Discourse Abingdon* 36, 647–664. doi: 10.1080/01596306.2014.930020

Rajchert, J. M., Żułtak, T., and Smulczyk, M. (2014). Predicting reading literacy and its improvement in the polish national extension of the PISA study: the role of intelligence, trait- and state-anxiety, socio-economic status and school-type. *Learn. Individ. Differ.* 33, 1–11. doi: 10.1016/j.lindif.2014.04.003

Rasmusson, M., and Åberg-Bengtsson, L. (2015). Does performance in digital reading relate to computer game playing? A study of factor structure and gender patterns in 15-year-olds' reading literacy performance. *Scand. J. Educ. Res.* 59, 691–709. doi: 10.1080/00313831.2014.965795

Richard, M. P. (2016). *The Dynamics of Persuasion: Communication and Attitudes in the Twenty-First Century*. New York, NY: Routledge.

Rutkowski, L., Gonzalez, E., Joncas, M., and Davier, M. V. (2010). International large scale assessment data issues in secondary analysis and reporting. *Educ. Res.* 39, 142–151. doi: 10.3102/0013189X10363170

Santos, M. A., Godas, A., Ferraces, M. J., and Lorenzo, M. (2016). Academic performance of native and immigrant students: a study focused on the perception of family support and control, school satisfaction, and learning environment. *Front. Psychol.* 7:1560. doi: 10.3389/fpsyg.2016.01560

Schachner, M. K., He, J., Heizmann, B., and Van de Vijver, F. J. R. (2017). Acculturation and school adjustment of immigrant youth in six european countries: findings from the programme for international student assessment (PISA). *Front. Psychol.* 8:649. doi: 10.3389/fpsyg.2017.00649

Sellar, S., and Lingard, B. (2013). The OECD and the expansion of PISA: new global modes of governance in education. *Brit. Educ. Res. J.* 40, 917–936. doi: 10.1002/berj.3120

Shin, S. H., Slater, C. L., and Backhoff, E. (2013). Principal perceptions and student achievement in reading in Korea, Mexico, and the united states: educational leadership, school autonomy, and use of test results. *Educ. Adm. Q.* 49, 489–527. doi: 10.1177/0013161X12458796

Skryabin, M., Zhang, J. J., Liu, L., and Zhang, D. (2015). How the ICT development level and usage influence student achievement in reading, mathematics, and science. *Comput. Educ.* 85, 49–58. doi: 10.1016/j.compedu.2015.02.004

Stekhoven, D. J., and Bühlmann, P. (2012). Missforest – non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 112–118. doi: 10.1093/bioinformatics/btr597

Stoet, G., and Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence* 48, 137–151. doi: 10.1016/j.intell.2014.11.006

Tay, L. Y., Nair, S. S., and Lim, C. P. (2017). A regression analysis of elementary students' ICT usage vis-à-vis access to technology in Singapore. *Educ. Media Int.* 54, 1–14. doi: 10.1080/09523987.2017.1324362

Torppa, M., Eklund, K., Sulkunen, S., Niemi, P., and Ahonen, T. (2018). Why do boys and girls perform differently on PISA reading in Finland? The effects of reading fluency, achievement behaviour, leisure reading and homework activity. *J. Res. Read.* 41, 122–139. doi: 10.1111/1467-9817.12103

Wei, Y., and Hu, J. (2018). A cross-sectional evaluation of EFL students' critical thinking dispositions in digital learning. *Adv. Soc. Sci. Educ. Hum. Res.* 195, 27–30. doi: 10.2991/iserss-18.2018.8

Wei, Y., Yang, Q., Chen, J., and Hu, J. (2018). The exploration of a machine learning approach for the assessment of learning styles changes. *Mechatron. Syst. Contr.* 46, 121–126. doi: 10.2316/Journal.201.2018.3.201-2979

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 48, 817–838. doi: 10.2307/1912934

Whyte, S., Schmid, E. C., Thompson, S. V. H., and Oberhofer, M. (2014). Open educational resources for call teacher education: the iTILT interactive whiteboard project. *Comput. Assist. Lang. Learn.* 27, 122–148. doi: 10.1080/09588221.2013.818558

Woessmann, L., and Fuchs, T. (2005). Computers and student learning: bivariate and multivariate evidence on the availability and use of computers at home and at school. *CESifo Working Pap. S.* 47, 339–347. doi: 10.1080/15391523.2005.10782441

Wooldridge, J. (2003). *Introductory Econometrics: A Modern Approach*, 5th Edn. Ohio: Thomson South-Western.

Xiao, Y., and Hu, J. (2019). Assessment of optimal pedagogical factors for Canadian ESL learners' reading literacy through artificial intelligence algorithms. *Int. J. Engl. Linguist.* 9, 1–14. doi: 10.5539/ijel.v9n4p1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

BRIEF RESEARCH REPORT
published: 30 July 2019
doi: 10.3389/fpsyg.2019.01742

Validation of the Measurement of Need Frustration

Isabeau K. Tindall[1]* *and Guy J. Curtis*[1,2]

[1] *Discipline of Psychology, Murdoch University, Murdoch, WA, Australia,* [2] *Discipline of Psychology, Murdoch University, and School of Psychological Science, The University of Western Australia, Perth, WA, Australia*

Until recently, need frustration was considered to be the absence of need satisfaction, rather than a separate dimension. Whilst the absence of need satisfaction can hamper growth, experiencing need frustration can lead to malfunctioning and subsequent psychopathology. Therefore, examining these constructs separately is vital, as they produce different outcomes, with the consequences of need frustration potentially more severe. This study sought to examine predictors of need frustration using undergraduate students and individuals from the wider community (N = 510, females N = 404, M_{age} = 24.15). Participants completed the new need satisfaction frustration scale and measures of anxiety, stress, depression, and negative and positive affect. Support for the position that need frustration is separate to Need Satisfaction and is related to psychological health problems (i.e., ill-being) was found. However, autonomy frustration was not found to be a significant predictor of ill-being. Extending previous research, this study found relationships of stress and somatic anxiety with need frustration. Further, a relationship between need frustration with anxiety and depression occurred, when these symptom dimensions were examined separately, through distinct questionnaires. Support for the construct of need frustration highlights the necessity of examining need frustration in addition to need satisfaction within future studies. Interventions specific to reducing need frustration, specifically competence and relatedness frustration within both the educational and workplace setting are outlined.

Keywords: need satisfaction, need frustration, NSFS, ill-being, anxiety, depression, stress, well-being

OPEN ACCESS

Edited by:
Elisa Pedroli,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Cristina Senín-Calderón,
University of Cádiz, Spain
Sandra Maria Correia Loureiro,
University Institute of Lisbon (ISCTE),
Portugal

****Correspondence:***
Isabeau K. Tindall
I.Tindall@Murdoch.edu.au

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *08 April 2019*
Accepted: *12 July 2019*
Published: *30 July 2019*

Citation:
Tindall IK and Curtis GJ (2019)
Validation of the Measurement
of Need Frustration.
Front. Psychol. 10:1742.
doi: 10.3389/fpsyg.2019.01742

INTRODUCTION

According to basic needs theory (BNT), a mini-theory of Self-Determination Theory (Deci and Ryan, 2000), individuals are motivated by three key psychological needs, the need for: autonomy, competence and relatedness. Autonomy is defined as the perception of control over one's behavior rather than feeling controlled by external factors. Competence relates to an individual's belief in their ability to attain desired outcomes, and Relatedness, the degree to which an individual feels closeness and a sense of belonging with others. BNT contends that meeting these needs are necessary for optimal human development (Ryan et al., 1996). The degree to which these needs are met, has direct implications to both the educational and workplace setting (education: Copeland and Levesque-Bristol, 2011; workplace: Gagné et al., 1997; Baard et al., 2004). Need satisfaction over these three domains, are strong predictors of first year retention rates, the perception of the university setting as a positive learning environment and increased academic performance (Copeland and Levesque-Bristol, 2011). Whilst, within the workplace, increased need satisfaction is related to better job performance, improved psychological well-being (Baard et al., 2004), increased work motivation (Gagné et al., 1997), and stronger employee commitment (Gagné et al., 2008). According to BNT the degree to which needs are met over the domains of autonomy, competency and relatedness, directly relate to a sense of subjective wellbeing, whilst having these needs frustrated, leads to ill-being. Therefore, according to this theory, need frustration is not a separate construct in and of itself, but rather, occurs because of the absence of need satisfaction (Deci and Ryan, 2000).

More recent research, however, has speculated that need frustration is not just the inverse of need satisfaction, but rather, is a distinct construct (Deci and Ryan, 2000; Sheldon and Gunz, 2009; Bartholomew et al., 2011a; Longo et al., 2016). Supporting this assertion, Longo et al. (2016) stated that these two constructs have separate theoretical underpinnings, and therefore, will predict different outcomes. Specifically, satisfaction related to the domains of autonomy, competence and relatedness, is associated with positive outcomes (well-being) such as positive affect. Whilst frustration related to these three domains (need frustration) can predict negative outcomes (ill-being) such as negative affect, depression and anxiety (Bartholomew et al., 2011a). Their study found support for this theory by showing significant relationships existed between satisfaction and positive outcomes, and frustration and negative outcomes. The study by Longo et al. (2016) is therefore, an extension of BNT (Deci and Ryan, 2000) due to the further refinement as to what defines, and predicts, satisfaction and frustration. Specifically, it was found that ill-being was uniquely predicted by experiences of need frustration and not from merely experiencing low need satisfaction (Longo et al., 2016). A recent study by Longo et al. (2018) found further support for the perspective that need satisfaction and need frustration are distinct constructs.

Due to the recency of the proposition that need frustration is separate from need satisfaction, limited research has been conducted into need frustration, and therefore, correlates of this construct. Research into need frustration is imperative; although a lack of need satisfaction is related to negative outcomes, increased need frustration is considered especially harmful, and linked to potential psychopathology (Bartholomew et al., 2011a). According to Vansteenkiste and Ryan (2013), this is illustrated through an example juxtaposing the repercussions of low need satisfaction as compared to experiencing increased need frustration within the workplace. An individual experiencing low need satisfaction through reduced relatedness with colleagues, might not feel as excited about their work. However, an individual actively bullied, ridiculed and excluded by colleagues, therefore, also experiencing low relatedness through high degrees of need frustration, will be at the additional risk of developing psychopathology, such as depression and severe stress. It can then be said that although a lack of need satisfaction can lead to a lack of fulfillment, need frustration, is also strongly related to malfunctioning (Vansteenkiste and Ryan, 2013). Given that the degree of mental illness has been steadily increasing within both the workplace (Bonde, 2008; Fan et al., 2015) and educational setting (American College Health Association, 2018), it is important to examine the potential link between need frustration and psychopathology.

In light of the above research, the present study aimed to further examine the relationship between psychological health problems, and need frustration. We did this through examining the relationship between the need satisfaction and frustration scale (NSFS) and ill-being found by Longo et al. (2016) through examining the factor structure of the NSFS. Particular attention was given to the predictors of need frustration due to the scarcity of research into this factor. Need satisfaction was not the focus of the present study as extensive research has been conducted into this construct (Deci and Ryan, 2002; Baard et al., 2004). Therefore, an omnibus of negative emotionality measures expected to be predicted by need frustration were included in this study. Negative emotionality measures of anxiety, stress, depression and negative affect, were included.

A limitation of the Longo et al. (2016) study, was the lack of ability to distinguish between anxiety and depression manifested through increased ill-being. Longo et al. (2016) used the General Health Questionnaire (GHQ, Goldberg and Williams, 1988) to measure the influence of ill-being on anxiety and depression, through the Anxiety-Depression subscale, which does not separate between these constructs. Although anxiety and depression share variance related to general distress, they are theoretically distinct dimensions, with anxiety uniquely related to social tension/arousal, and depression; anhedonia/low affect (Clark and Watson, 1991). Therefore, to allow for an examination of the distinct relationship between anxiety and depression with ill-being, we included well validated measures of both anxiety and depression. To allow for replication of Longo et al. (2016), a measure of negative affect was also included. Although an investigation into need frustration has already occurred within the sporting context (Bartholomew et al., 2011b), and within general life (Sheldon and Gunz, 2009), the present study sought to directly extend on the study by Longo et al. (2016). Therefore, we examined the influence of need frustration on negative emotionality specifically within the context of the educational setting and workplace. According to research into the educational setting (Riolli et al., 2012), university students experience high levels of stress. Further, the workplace can be inherently stressful (Colligan and Higgins, 2006). Therefore, in extension of Longo et al. (2016), in addition to measures of negative affect, depression and anxiety, we also included a measure of stress.

In light of the above aim, it was hypothesized that:

(1) Measures of psychological health problems, specifically those measuring depression, anxiety, stress and negative affect, will negatively relate to satisfaction and positively relate to frustration.
(2) Need Satisfaction will be positively correlated with positive affect.
(3) That measures related to psychological health problems and positive affect will be strongly related to their proposed factors, highlighting support for the factor structure put forward by Longo et al. (2016).

MATERIALS AND METHODS

Participants

A sample of 510 (females N = 404, M_{age} = 24.15, SD = 8.06; range = 18–59), undergraduate students from Murdoch University and members of the wider community (78% Caucasian) participated in this study for partial course credit, or the potential to gain a gift voucher, respectively. Ethics approval was acquired from Murdoch University before data collection.

Materials

Need satisfaction and frustration scale (NSFS; Longo et al., 2016). The NSFS consists of six 3-item subscales, measuring need satisfaction in the domains of autonomy, relatedness and competence and the other three, measuring need frustration in these domains. This scale examined these needs in the context of work and/or educational settings. Items are rated on a 7-point Likert scale from 1 (*strongly disagree*), to 7 (*strongly agree*), higher scores on the satisfaction subscales indicate greater need satisfaction, whilst higher scores on the frustration subscales indicate increased frustration. An example item is "In my studies/In my job... I feel, I'm given a lot of freedom in deciding how I do things." The subscales of autonomy, relatedness, and competence over the domains of satisfaction and frustration have exhibited excellent reliability (αs > 0.70) in both the educational and workplace context. Further, the criterion validity of this scale with other measures of need satisfaction are sufficient over both settings ($rs \geq 0.4$; Longo et al., 2016).

Omnibus affect measures. Anxiety was measured through the State-Trait Inventory for Cognitive and Somatic Anxiety (STICSA; Ree et al., 2008), the State-Trait Anxiety Inventory (STAI; Spielberger et al., 1970, 1983), the Anxiety Sensitivity Index (ASI; Reiss et al., 1986) and the anxiety subscale of the Depression Anxiety Stress Scale-21 (DASS-21; Lovibond and Lovibond, 1995). All anxiety measures exhibited sound validity and internal consistency previously (αs > 0.83; Spielberger et al., 1983; Peterson and Heilbronner, 1987; Lovibond and Lovibond, 1995; Grös et al., 2007). Depression was measured through the depression subscale of the DASS-21 (Lovibond and Lovibond, 1995) and the Beck Depression Inventory-II (BDI-II; Beck et al., 1996). For the BDI-II, item 9, "Suicidal Thoughts or Wishes" was removed according to a requirement by the Ethics Committee of Murdoch University. Depression measures included have previously exhibited good validity and internal consistency (αs > 0.84; Lovibond and Lovibond, 1995; Dozois et al., 1998). Stress was measured through the stress subscale of the DASS-21 (Lovibond and Lovibond, 1995). This subscale has also exhibited sound internal consistency and reliability (α = 0.90; Lovibond and Lovibond, 1995). Positive affect and negative affect were measured using an adapted version of the Positive and Negative Affect Schedule-X (Watson and Clark, 1994; Church et al., 2014).

Both subscales of positive and negative affect exhibit strong previous internal consistency and reliability (α = 0.83; Watson and Clark, 1994). Response scales for these measures were as they appear in their original sources or manuals.

Procedure

Participants gave written informed consent and completed questionnaires online. Participants were also told their responses to these questionnaires would be anonymous. The order of questionnaires presented were randomized. Students were recruited through a participant database at Murdoch University and through fliers posted around the university, whilst community members were recruited through social media. The surveys took approximately 30 min to complete.

RESULTS

Data was non-normal, however, a large sample size was used, and so normality was assumed (Ghasemi and Zahediasl, 2012). Little's (1988) MCAR test was non-significant and missing values consisted of <5% of the total sample, therefore missing values were imputed with the series mean. Using a Z of ±3.29 for assessing outliers (Field, 2009), responses from seven participants were removed. A total of 503 participants were therefore included in the final analysis.

Participants also completed the trait versions of the STICSA cognitive and somatic subscales and the STAI trait, however, these were not included in the analysis of ill-being, as need frustration only theoretically affects state anxiety, as trait anxiety should be stable over time (Spielberger et al., 1983; Ree et al., 2008). Due to overlapping symptom dimensions of the anxiety, depression, stress and negative and positive affect measures, multicollinearity of these measures was checked. No measures exceeded multicollinearity cut offs according to values of the VIF < 10 and tolerance > 0.1 (Hair et al., 1995).

Correlations between the NSFS questionnaire subscales and the measures of interest, as well as descriptive statistics and reliability estimates are reported in **Table 1**.

As seen in **Table 1**, internal consistencies for all measures were good, with all alphas >0.7 (Cronbach, 1951). All survey items measuring ill-being outcomes were positively correlated with frustration scales on the NSFS. The PANAS-Positive was positively correlated with satisfaction scales.

A structural equation model (SEM) was then calculated in AMOS 24 using a maximum likelihood estimation procedure, to assess the factor structure of the NSFS and alignment with the measures of interest (see **Figure 1**).

Results of the non-constrained model suggested an unacceptable fit: $\chi^2(84)$ = 534.215, p < 0.001. CFI = 0. 908, TLI = 0.869, GFI = 0.880, RMR = 3.479, RMSEA = 0.103 (90% CI = 0.095−0.112) and CMIN/DF = 6.360. However, modification indices suggested freeing error variance between the error terms of some of the anxiety, stress, negative affect, and depression measures loading onto ill-being. With regard to correlating the error terms, only measures with strong theoretical support for association were correlated (Cole et al., 2007; Hooper et al., 2008). Subsequently, only errors from measures of anxiety and stress were allowed to correlate (Lovibond and Lovibond, 1995; Roberts et al., 2016), whilst errors from negative affect was only allowed to correlate with depression (Danhauer et al., 2013). For the constrained model, the chi-square value for the

TABLE 1 | Correlations between the six subscales of the NSFS, ill-being, and the PANAS-Positive.

			Measures									
			PANAS positive	PANAS negative	BDI-II	ASI	DASS anxiety	DASS stress	DASS depression	STICSA state somatic	STICSA state cognitive	STAI state
		***M* (SD)**	**32.53 (7.12)**	**22.08 (7.61)**	**13.80 (9.77)**	**22.96 (11.98)**	**9.67 (8.73)**	**14.75 (9.48)**	**10.86 (9.91)**	**15.44 (5.03)**	**19.22 (7.05)**	**41.66 (12.09)**
Sub-scale	***M* (SD)**	α	**0.90**	**0.90**	**0.90**	**0.90**	**0.84**	**0.85**	**0.91**	**0.87**	**0.90**	**0.94**
NSFS autonomy satisfaction	14.38 (4.06)	0.87	0.29**	−0.27**	−0.25**	−0.20**	−0.19**	−0.20**	−0.21**	−0.22**	−0.21**	−0.31**
NSFS autonomy frustration	11.60 (3.75)	0.75	−0.24**	0.27**	0.26**	0.25**	0.18**	0.26**	0.22**	0.15**	0.24**	0.27**
NSFS relatedness satisfaction	13.36 (3.78)	0.80	0.44**	−0.39**	−0.39**	−0.21**	−0.25**	−0.27**	−0.40**	−0.18**	−0.28**	−0.32**
NSFS relatedness frustration	12.03 (3.92)	0.76	−0.44**	0.51**	0.47**	0.37**	0.38**	0.41**	0.45**	0.26**	0.43**	0.42**
NSFS competence satisfaction	13.81 (3.54)	0.86	0.53**	−0.43**	−0.44**	−0.26**	−0.30**	−0.30**	−0.48**	−0.26**	−0.42**	−0.45**
NSFS competence frustration	12.44 (4.01)	0.80	−0.46**	0.45**	0.46**	0.35**	0.32**	0.40**	0.48**	0.24**	0.44**	0.44**

*$p < 0.05$, **$p < 0.01$.

overall model fit was significant, $\chi^2(77) = 275.996$, $p < 0.001$, suggesting a lack of fit between the hypothesized model and the data. However, due to the oversensitivity of the χ^2 to large sample sizes, other fit indices were assessed (Kline, 1998). Examination of these other indices showed acceptable model fit (Hu and Bentler, 1999; Longley et al., 2005) with CFI = 0.959, TLI = 0.937, GFI = 0.941, RMR = 2.675, CMIN/DF = 3.584 and RMSEA = 0.072 (90% CI = 0.063−0.081).

For frustration items, competence and relatedness frustration loaded significantly onto ill-being. Further, satisfaction items related to competence and relatedness satisfaction loaded significantly onto the PANAS-positive. Autonomy frustration did not load significantly onto ill-being, nor did autonomy satisfaction significantly load onto positive affect. All measures examining negative outcomes loaded significantly onto the latent ill-being factor.

DISCUSSION

The present study aimed to extend on limited research into need frustration within the educational and workplace setting. We also sought to further justify the separation between need frustration and satisfaction put forward by Longo et al. (2016). Further, due to the prevalence of mental illness within the educational (American College Health Association, 2018) and workplace setting (Bonde, 2008; Fan et al., 2015), we examined the relationship between psychological health problems and need frustration.

In support of hypothesis one, positive correlations between the measures assessing ill-being and frustration occurred. Further, the negative correlation with these measures and satisfaction increases support for this hypothesis. Measures of negative affect, depression, stress and most measures of anxiety, were also moderately positively correlated with relatedness and competence frustration. These measures were similarly negatively associated with relatedness and competence satisfaction. In most cases, ill-being measures were only weakly correlated with autonomy satisfaction and frustration. This suggests that psychological health problems might not be strongly related to feelings of autonomy. With regard to the educational setting, this is plausible, considering that autonomy has been found to be a weak predictor of positive outcomes, such as academic motivation within undergraduate populations (Grolnick et al., 2002; Faye and Sharpe, 2008).

Hypothesis two was also supported as a positive relationship was found between positive affect and satisfaction, whereas this relationship was negative with frustration. Like the measures of ill-being, autonomy satisfaction had the weakest relationship with positive affect. Future research should seek to extend the examination of autonomy beyond the undergraduate population, through testing post-graduate students. Within the workplace, a need for autonomy is a critical predictor of well-being, performance, motivation, and reduced emotional distress (Gagné and Bhave, 2011). Post-graduate students are expected to feel a stronger need for autonomy than undergraduate students, through increased control over output, therefore the satisfaction

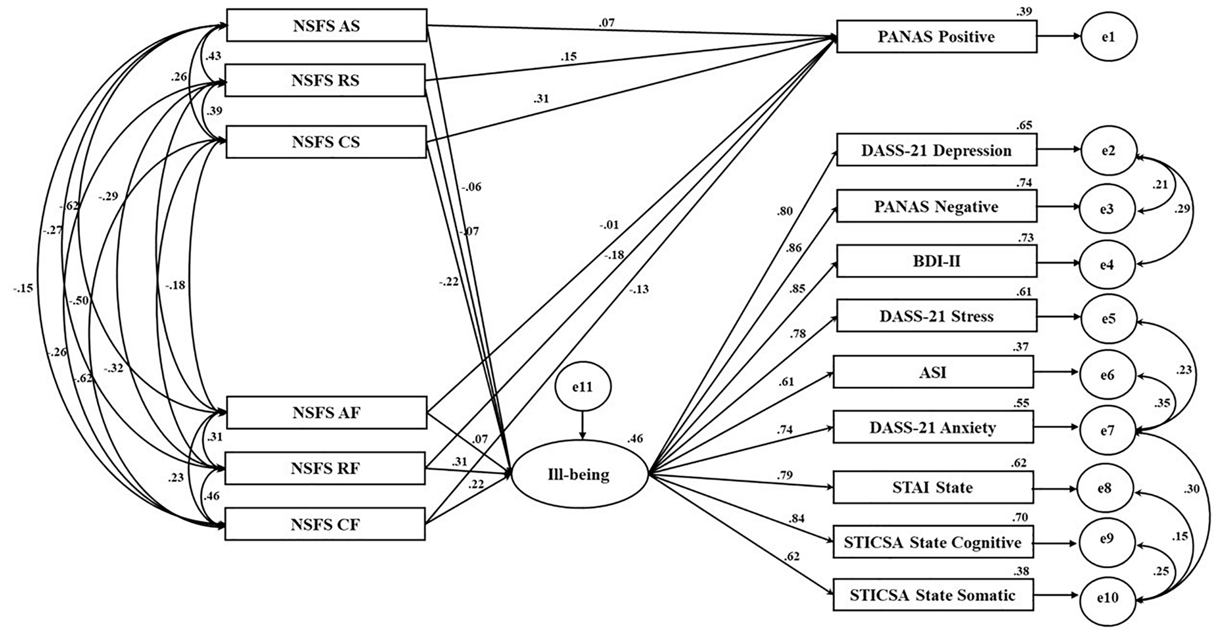


FIGURE 1 | Predicting ill-being outcomes and relationship between satisfaction (NSFS) and positive affect (PANAS-Positive). *A*, autonomy; *S*, satisfaction; *R*, relatedness; *C*, competence; *F*, frustration; *e*, error.

and frustration of autonomy might be more important to this population.

Hypothesis three was partially supported. Relatedness satisfaction and competence satisfaction significantly loaded onto positive affect, and relatedness frustration and competence frustration loaded significantly onto ill-being. However, autonomy satisfaction did not significantly load onto positive affect, and autonomy frustration did not load significantly onto ill-being.

In the original study conducted by Longo et al. (2016), additional measures of vigor, intrinsic motivation and job satisfaction, were examined in relation to their relationship with wellbeing. This study did not include these measures and therefore, only examined the relationship between positive affect and need satisfaction. This could have reduced potential factor loadings between autonomy satisfaction and positive affect, as autonomy satisfaction might be more strongly related to these additional measures. Indeed, the factor loading for autonomy satisfaction and well-being in the study by Longo et al. (2016) was higher and significant when these other measures were included. A similar proposition can explain the non-significant loading between autonomy frustration and ill-being. Despite this study extensively examining the relationship between ill-being outcomes, we did not include a measure assessing exhaustion. Therefore, autonomy frustration might be strongly related to exhaustion. Measures of exhaustion, intrinsic motivation, vigor and job satisfaction, should be included when examining the relationship with well-being and ill-being during future studies.

The finding that state somatic anxiety as measured by the STICSA significantly loaded onto ill-being and that this was related to need frustration, is important theoretically. It suggests that need frustration is related to physiological anxiety symptoms such as automatic nervous system arousal (Ree et al., 2008). This supports Bartholomew et al. (2011b) who found that need thwarting (frustration) was related to somatic complaints in athletes. This is in extension of Longo et al.'s (2016) findings as they did not examine the relationship between the NSFS and somatic anxiety. It also elaborates on Bartholomew et al. as somatisation was found to influence need frustration outside the domain of athletes. Further, the significant relationship between ill-being and stress as measured by the DASS-Stress scale is also theoretically important, as this finding is novel, and therefore in extension of previous research (Longo et al., 2016).

Relatedness frustration had the strongest relationship with ill-being. This supports previous research (Larson et al., 1996). Negative affect and depression strongly, and significantly, loaded onto ill-being. Relatedness frustration directly relates to experiencing social exclusion and loneliness (Chen et al., 2015), with prolonged periods of loneliness associated with depression and increased negative affect (van Winkel et al., 2017). Interventions within the university setting should focus on instilling a greater sense of inclusion within university students, which in turn, might reduce symptoms of depression and negative affect. In the study conducted by Mattanah et al. (2010), students that partook in a peer-led social inclusion program, experienced increased feelings of social inclusion and reduced loneliness.

In addition to interventions specific to improving peer relationships, research has also highlighted to importance of the teacher in terms of fostering a sense of inclusion, with this subsequently, strengthening relatedness. Students reported increased relatedness when they felt that their teacher genuinely cared, respected and valued them (Niemiec and Ryan, 2009). Therefore, lecturers and tutors should endeavor to convey increased warmth, caring and respect toward students (Niemiec and Ryan, 2009). This finding also has implications for the workplace, as it has been found that transformational leaders, who foster relatedness, through increased employee respect, and through instilling a sense of cohesion through shared team goals, improved outcomes (Kovjanic et al., 2013). Therefore, transformational leadership training programs (Hasson et al., 2016) focusing on improving leader/followers' relationships should be implemented (Dvir et al., 2002).

State cognitive anxiety as measured by the STICSA, was also strongly related to ill-being. After relatedness frustration, competence frustration was the strongest predictor of ill-being. Competence frustration relates to negative feelings an individual has toward their self-efficacy and increased feelings of failure (Sweet et al., 2012; Chen et al., 2015). Like with depression, increased anxiety is associated with low self-efficacy (Jerusalem and Schwarzer, 1992). Therefore, in addition to relatedness frustration increasing the manifestation of psychological health problems, competence frustration might also contribute to ill-being. Therefore, interventions within both the educational and workplace setting, should also target competence. According to Niemiec and Ryan (2009) competence within the educational setting can be improved through rewarding effort, in addition to academic performance. Some students report that despite immense effort, they do not receive the academic performance they expect, and therefore feel their effort has been under-rewarded (Copeland and Levesque-Bristol, 2011). This feeling of inadequacy consequently reduces competence. Presently, most scholarships within the university context are awarded based on academic merit. Despite grades reflecting motivation to learn in some cases, for some students, effort is more indicative of performance. Therefore, to instill a feeling of competence, some scholarships could be awarded to students based on the level of effort or degree of improvement a student makes (Copeland and Levesque-Bristol, 2011).

This study lends support to the proposition that need satisfaction and need frustration are separate constructs (Longo et al., 2016, 2018). Further, the finding that need frustration is strongly related to psychological health problems, specifically negative affect, depression, anxiety and stress, extends research into need frustration (Longo et al., 2016). Unlike Longo et al. (2016), this study separately measured the manifestation of state anxiety and depression symptoms created through increased need frustration. This study also examined the influence of need frustration on the expression of stress. The current study highlights the magnitude of potential ill-being outcomes created through increased frustration of psychological needs, specifically competence and relatedness frustration, expressed within the workplace and/or educational settings.

Limitations and Future Directions

Potential limitations of this study are the lack of measures examining vigor and intrinsic motivation. Despite this study's main aim seeking to extensively examine the predictors of ill-being, future research should include more well-being measures. This will allow for a deeper examination into the claim that need frustration and need satisfaction are distinct constructs instead of need frustration relating to the absence of need satisfaction (Longo et al., 2016, 2018). Further, a measure of exhaustion should be included to examine whether autonomy frustration is a significant predictor of Need Frustration, or if its inclusion in the NSFS should be reviewed.

Future research should seek to implement the suggested interventions related to reducing competence and relatedness frustration within both the workplace and university setting. Within the educational setting, it was recommended that depression and negative affect could be reduced through peer-led social inclusion programs fostering social inclusion and reducing isolation (Mattanah et al., 2010). Rewarding effort, rather than academic merit, via the implementation of effort-based scholarships (Copeland and Levesque-Bristol, 2011) might increasing self-efficacy and competence, subsequently decreasing anxiety and depression. Lastly, within both the educational and workplace setting, lecturers, tutors and leaders, could more outwardly express respect and value toward their students and employees to improve relatedness. To quantifying the magnitude of improvement once interventions are implemented, a longitudinal design should be used. Within the university setting, psychological need satisfaction/frustration could be measured when students first start university, to act a baseline, and measured once again after the implementation of interventions. To quantify the retention of positive outcomes after implementation, additional measures should be taken for the duration of the student's undergraduate degree.

CONCLUSION

The current study gives preliminary support to Longo et al. (2016, 2018), who stated that need frustration and need satisfaction are distinct constructs. Theoretically, this study also gives further insight into the relationship between basic need frustration and common types of psychological health problems, such as anxiety specific to physiological symptoms, and stress. Whilst practically, potential interventions to reduce need frustration and reduce psychological symptoms of ill-being are presented.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the National Statement on Ethical Conduct in Human Research, 2007, National Health and Medical Research Council Act 1992, with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Human Research Ethics Committee at Murdoch University.

AUTHOR CONTRIBUTIONS

IT collected and analyzed the data, and prepared the draft and final manuscript. GC provided feedback on draft manuscripts to prepare it for publication.

FUNDING

IT was supported by an Australian Commonwealth Government Research Training Scheme Scholarship.

ACKNOWLEDGMENTS

We would like to thank all the participants who participated and completed the questionnaires.

REFERENCES

American College Health Association (2018). *American College Health Association-National College Health Assessment II: Undergraduate Student Reference Group Executive Summary Spring 2018*. Hanover: American College Health Association, 2018.

Baard, P. P., Deci, E. L., and Ryan, R. M. (2004). Intrinsic need satisfaction: a motivational basis of performance and weil ∼being in two work settings. *J. Appl. Soc. Psychol.* 34, 2045–2068. doi: 10.1111/j.1559-1816.2004.tb02690.x

Bartholomew, K. J., Ntoumanis, N., Ryan, R. M., Bosch, J. A., and Thøgersen-Ntoumani, C. (2011a). Self-determination theory and diminished functioning: the role of interpersonal control and psychological need thwarting. *Personal. Soc. Psychol. Bull.* 37, 1459–1473. doi: 10.1177/0146167211413125

Bartholomew, K., Ntoumanis, N., Ryan, R. M., and Thøgersen-Ntoumani, C. (2011b). Psychological need thwarting in the sport context: assessing the darker side of athletic experience. *J. Sport Exerc. Psychol.* 33, 75–102. doi: 10.1123/jsep.33.1.75

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). *Manual for the Beck Depression Inventory-*, Vol. II. San Antonio, TX: Psychological Corporation, 78.

Bonde, J. P. E. (2008). Psychosocial factors at work and risk of depression: a systematic review of the epidemiological evidence. *Occup. Environ. Med.* 65, 438–445. doi: 10.1136/oem.2007.038430

Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E. L., Van der Kaap-Deeder, J., et al. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motiv. Emot.* 39, 216–236. doi: 10.1007/s11031-014-9450-1

Church, A. T., Katigbak, M. S., Ibáñez-Reyes, J., de Jesús Vargas-Flores, J., Curtis, G. J., Tanaka-Matsumi, J., et al. (2014). Relating self-concept consistency to hedonic and eudaimonic well-being in eight cultures. *J. Cross Cult. Psychol.* 45, 695–712. doi: 10.1177/0022022114527347

Clark, L. A., and Watson, D. (1991). Tripartite model of anxiety and depression: psychometric evidence and taxonomic implications. *J. Abnorm. Psychol.* 100, 316–336. doi: 10.1037/0021-843X.100.3.316

Cole, D. A., Ciesla, J. A., and Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychol. Methods* 12, 381–398. doi: 10.1037/1082-989X.12.4.381

Colligan, T. W., and Higgins, E. M. (2006). Workplace stress: etiology and consequences. *J. Workplace Behav. Health* 21, 89–97. doi: 10.1300/J490v21n02_07

Copeland, K. J., and Levesque-Bristol, C. (2011). The retention dilemma: effectively reaching the first-year university student. *J. Coll. Stud. Ret.* 12, 485–515. doi: 10.2190/CS.12.4.f

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/bf02310555

Danhauer, S. C., Legault, C., Bandos, H., Kidwell, K., Costantino, J., Vaughan, L., et al. (2013). Positive and negative affect, depression, and cognitive processes in the cognition in the Study of tamoxifen and raloxifene (Co-STAR) Trial. *Aging Neuropsychol. Cogn.* 20, 532–552. doi: 10.1080/13825585.2012.747671

Deci, E. L., and Ryan, R. M. (2000). The "what" and "why" of goal pursuits: human needs and the self-determination of behavior. *Psychol. Inquiry* 11, 227–268. doi: 10.1207/S15327965PLI1104_01

Deci, E, L., and Ryan, R. M. (eds). (2002). *Handbook of Self-Determination Research*. Rochester, NY: University of Rochester Press.

Dozois, D. J., Dobson, K. S., and Ahnberg, J. L. (1998). A psychometric evaluation of the beck depression inventory–II. *Psychol. Assess.* 10, 83–89. doi: 10.1037/1040-3590.10.2.83

Dvir, T., Eden, D., Avolio, B. J., and Shamir, B. (2002). Impact of transformational leadership on follower development and performance: a field experiment. *Acad. Manag. J.* 45, 735–744. doi: 10.5465/3069307

Fan, L. B., Blumenthal, J. A., Watkins, L. L., and Sherwood, A. (2015). Work and home stress: associations with anxiety and depression symptoms. *Occup. Med.* 65, 110–116. doi: 10.1093/occmed/kqu181

Faye, C., and Sharpe, D. (2008). Academic motivation in university: the role of basic psychological needs and identity formation. *Can. J. Behav. Sci.* 40, 189–199. doi: 10.1037/a0012858

Field, A. (2009). *Discovering Statistics Using SPSS*, 3rd Edn. London: Sage Publications Ltd.

Gagné, M., and Bhave, D. (2011). *"Autonomy in the Workplace: An Essential Ingredient to Employee Engagement and Well-being in Every Culture". In Human autonomy in cross-cultural context*. Dordrecht: Springer, 163–187.

Gagné, M., Chemolli, E., Forest, J., and Koestner, R. (2008). A temporal analysis of the relation between organisational commitment and work motivation. *Psychol. Belg.* 48, 219–241. doi: 10.5334/pb-48-2-3-219

Gagné, M., Senécal, C. B., and Koestner, R. (1997). Proximal job characteristics. Feelings of empowerment, and intrinsic motivation: a multidimensional model. *J. Appl. Soc. Psychol.* 21, 1222–1240. doi: 10.1111/j.1559-1816.1997.tb01803.x

Ghasemi, A., and Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *Int. J. Endocrinol. Metab.* 10, 486–489. doi: 10.5812/ijem.3505

Goldberg, D., and Williams, P. D. P. M. (1988). *A User's Guide to the General Health Questionnaire. 1988*. Windsor: NFER-Nelson.

Grolnick, W. S., Gurland, S. T., DeCourcey, W., and Jacob, K. (2002). Antecedents and consequences of mothers' autonomy support: an experimental investigation. *Dev. Psychol.* 38, 143–155. doi: 10.1037/0012-1649.38.1.143

Grös, D. F., Antony, M. M., Simms, L. J., and McCabe, R. E. (2007). Psychometric properties of the state-trait inventory for cognitive and somatic anxiety (sticsa): comparison to the state-trait anxiety inventory (STAI). *Psychol. Assess.* 19, 369–381. doi: 10.1037/1040-3590.19.4.369

Hair, J. F., Anderson, R. E., Tatham, R. L., and Black, W. C. (1995). *Multivariate Data Analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Hasson, H., von Thiele Schwarz, U., Holmstrom, S., Karanika-Murray, M., and Tafvelin, S. (2016). Improving organizational learning through leadership training. *J. Workplace Learn.* 28, 115–129. doi: 10.1108/JWL-06-2015-0049

Hooper, D., Coughlan, J., and Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *Electron. J. Buis. Res. Methids* 6, 53–60. doi: 10.1016/j.acap.2015.07.001

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives, structural equation modeling: a Multidisciplinary. *Journal* 6, 1–55. doi: 10.1080/10705519909540118

Jerusalem, M., and Schwarzer, R. (1992). “"Self-efficacy as a resource factor in stress appraisal processes",” in *Self-efficacy: Thought Control of Action*, ed. R. Schwarzer (Washington DC: Hemisphere), 195–213.

Kline, R. B. (1998). *Principle and Practice of Structural Equation Modeling*. New York, NY: Guilford Publications.

Kovjanic, S., Schuh, S. C., and Jonas, K. (2013). Transformational leadership and performance: an experimental investigation of the mediating effects of basic needs satisfaction and work engagement. *J. Occup. Organ. Psychol.* 86, 543–555. doi: 10.1111/joop.12022

Larson, R. W., Richards, M. H., Moneta, G., Holmbeck, G., and Duckett, E. (1996). Changes in adolescents' daily interactions with their families from ages 10 to 18: disengagement and transformation. *Dev. Psychol.* 32, 744–754. doi: 10.1037/0012-1649.32.4.744

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *J. Am. Statist. Assoct.* 83, 1198–1202. doi: 10.1080/01621459.1988.10478722

Longley, S. L., Watson, D., and Noyes, R. Jr. (2005). Assessment of hypochondriasis domain: the multidimensional inventory of hypochondriacal traits (MIHT). *Psychol. Assess.* 17, 3–14. doi: 10.1037/1040-3590.17.1.3

Longo, Y., Alcaraz-Ibáñez, M., and Sicilia, A. (2018). Evidence supporting need satisfaction and frustration as two distinguishable constructs. *Psicothema* 30, 74–81. doi: 10.7334/psicothema2016.367

Longo, Y., Gunz, A., Curtis, G. J., and Farsides, T. (2016). Measuring need satisfaction and frustration in educational and work contexts: the need satisfaction and frustration scale (NSFS). *J. Happiness. Stud.* 17, 295–317. doi: 10.1007/s10902-014-9595-3

Lovibond, P. F., and Lovibond, S. H. (1995). The structure of negative emotional states: comparison of the depression anxiety stress scales (DASS) with the beck depression and anxiety inventories. *Behav. Res. Ther.* 33, 335–343. doi: 10.1016/0005-7967(94)00075-U

Mattanah, J. F., Ayers, J. F., Brand, B. L., Brooks, L. J., Quimby, J. L., and McNary, S. W. (2010). A social support intervention to ease the college transition: exploring main effects and moderators. *J. Coll. Stud. Deve.* 51, 93–108. doi: 10.1353/csd.0.0116

Niemiec, C. P., and Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: applying self-determination theory to educational practice. *School Field* 7, 133–144. doi: 10.1177/1477878509104318

Peterson, R. A., and Heilbronner, R. L. (1987). The anxiety sensitivity index:: construct validity and factor analytic structure. *J. Anxiety Disord.* 1, 117–121. doi: 10.1016/0887-6185(87)90002-8

Ree, M. J., French, D., MacLeod, C., and Locke, V. (2008). Distinguishing cognitive and somatic dimensions of state and trait anxiety: development and validation of the state-trait inventory for cognitive and somatic anxiety (STICSA). *Behav. Cogn. Psychother*. 36, 313–332. doi: 10.1017/S1352465808004232

Reiss, S., Peterson, R. A., Gursky, D. M., and McNally, R. J. (1986). Anxiety sensitivity, anxiety frequency and the prediction of fearfulness. *Behav. Res. Ther.* 24, 1–8. doi: 10.1016/0005-7967(86)90143-9

Riolli, L., Savicki, V., and Richards, J. (2012). Psychological capital as a buffer to student stress. *Psychology* 3, 1202–1207. doi: 10.4236/psych.2012.312A178

Roberts, K. E., Hart, T. A., and Eastwood, J. D. (2016). Factor structure and validity of the state-trait inventory for cognitive and somatic anxiety. *Psychol. Assess.* 28, 134–146. doi: 10.1037/pas0000155

Ryan, R. M., Sheldon, K. M., Kasser, T., and Deci, E. L. (1996). "All goals are not created equal: an organismic perspective on the nature of goals and their regulation," in *The psychology of action: Liking cognitivion to Behaviour*, eds P. M. Gollwitzer and J. A. Bargh (New York, N.Y: Guildford Press), 7–26.

Sheldon, K. M., and Gunz, A. (2009). Psychological needs as basic motives, not just experiential requirements. *J. Pers.* 77, 1467–1492. doi: 10.1111/j.1467-6494.2009.00589.x

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., and Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C. D., Gorsuch, R. L., and Lushene, R. E. (1970). *STAI: Manual for the STATE-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Sweet, S. N., Fortier, M. S., Strachan, S. M., and Blanchard, C. M. (2012). Testing and integrating self-determination theory and self-efficacy theory in a physical activity context. *Can. Psychol. Psychol. Can.* 53, 319–327. doi: 10.1037/a0030280

van Winkel, M., Wichers, M., Collip, D., Jacobs, N., Derom, C., Thiery, E., et al. (2017). Unraveling the role of loneliness in depression: the relationship between daily life experience and behavior. *Psychiatry* 80, 104–117. doi: 10.1080/00332747.2016.1256143

Vansteenkiste, M., and Ryan, R. M. (2013). On psychological growth and vulnerability: basic psychological need satisfaction and need frustration as a unifying principle. *J. Psychother. Int.* 23, 263–280. doi: 10.1037/a0032359

Watson, D., and Clark, L. A. (1994). *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form*. Ames: The University of Iowa.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 02 August 2019
doi: 10.3389/fpsyg.2019.01770

Development and Validation of a Pioneer Scale on Service Leadership Behavior in the Service Economies

Daniel T. L. Shek[1]*, *Diya Dou*[1] *and Lawrence K. Ma*[2]

[1] Department of Applied Social Sciences, The Hong Kong Polytechnic University, Kowloon, Hong Kong, [2] Department of Psychology, The Education University of Hong Kong, Tai Po, Hong Kong

In response to the severe lack of leadership assessment tools in the Chinese context, the Service Leadership Behavior Scale was developed based on the Service Leadership Model proposed by Po Chung, the co-founder of DHL International. Utilizing responses from 4,486 Hong Kong undergraduates, this paper reports the findings of a validation study on the Short-Form Service Leadership Behavior Scale (SLB-SF-65). Previous findings based on exploratory factor analysis supported a six-factor 48-item solution (SLB-SF-48). With the removal of ten items, confirmatory factor analysis showed that the final 38-item scale (SLB-SF-38) possessed excellent internal consistency, concurrent validity, and factorial validity based on multigroup invariance analyses. Overall speaking, the present study underscores the utility of the SLB-SF-38 as an objective assessment instrument of service leadership behavior in the education, research and personnel training contexts.

Keywords: scale validation, service leadership, leadership education, confirmatory factor analysis, Hong Kong

OPEN ACCESS

Edited by:
Elisa Pedroli,
Italian Auxological Institute, Italy

Reviewed by:
Xiongzhao Zhu,
Central South University, China
Andrea Bonanomi,
Catholic University of the Sacred Heart, Italy

****Correspondence:***
Daniel T. L. Shek
daniel.shek@polyu.edu.hk

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *07 April 2019*
Accepted: *16 July 2019*
Published: *02 August 2019*

Citation:
Shek DTL, Dou D and Ma LK (2019) Development and Validation of a Pioneer Scale on Service Leadership Behavior in the Service Economies. Front. Psychol. 10:1770. doi: 10.3389/fpsyg.2019.01770

INTRODUCTION

Over the past few decades, a structural transformation from the manufacturing-based to service-focused economies has been observed in many developed as well as developing countries (Bryson and Daniels, 2015; Snell et al., 2017). As such, possessing effective leadership qualities in this service era is indispensable in the contemporary world (Chung, 2015; Chung and Elfassy, 2016).

This service-focused leadership has been widely discussed in literature on both public and commercial service units. According to Schneider et al. (2005), leader's service-focused behavior, or service leadership, communicates a commitment to high levels of service quality. Compared with general leadership, service leadership is believed to exert a stronger influence on service outcomes (Hong et al., 2013). It is argued that service-oriented management and effective service leadership foster a service climate and consequently improve service performance (Jiang et al., 2015). Some assessment tools on service leadership have been developed and adopted in related empirical studies (Schneider et al., 2005; Jiang et al., 2015), such as Service Climate Scale (includes items measuring service-oriented leadership behavior) developed by Schneider et al. (1998), and a managerial measure of organizational service-orientation developed by Lytle et al. (1998), where service leadership was conceptualized as a combination of servant leadership and service orientation.

Although available scales measuring service leadership have a solid theoretical foundation and engendered much research, some research gaps exist. First, these scales were often developed with a strong focus on customer service. However, "service" in service economy should be interpreted in a broader context involving not only customer service but also the commitment to self-development, service to followers as well as society. Second, although service leadership is closely related to servant leadership, they are distinct concepts (Sendjaya and Sarros, 2002; Wong et al., 2015).

According to the servant leadership theory, followers' needs precede leaders' individual needs (Shek et al., 2015a). In contrast, service leadership seeks the mutual satisfaction of needs of both leaders and followers. Therefore, servant leadership scales may not be totally appropriate to assess service leadership. Third, available scales of service leadership mainly focus on leadership competences that guide and reward service delivery (i.e., "doing" of service leadership), such as goal setting, planning and coordinating (Schneider et al., 2005). Leaders' ability to make moral decisions and caring for others (i.e., "being" of service leadership) have often been considered relevant factors but not indispensable attributes of service leadership (Jiang et al., 2016). To fill the gaps, a set of assessment tools measuring service leadership was developed based on the Service Leadership Model proposed by Po Chung (Shek et al., 2015b, 2018a). In the following parts, the Service Leadership Model, its unique features, and the project entailing the construction and validation of Service Leadership Scales are outlined.

The Service Leadership Model and Its Unique Features

Service leadership is conceptualized as a "service aimed at ethically satisfying the need of self, others, groups, communities, systems, and environments" (Shek and Lin, 2015a, p. 233). The Service Leadership Model highlights three core attributes: *Competence*, *Character*, and *Caring*. First, *Competence* covers one's task-specific knowledge and skill sets required to excel in operational duties, which are essential for leaders to win over their followers (Chung and Bell, 2015). *Character* is defined as one's propensity to behave "in ways that are consistent with high [moral] values" (Chung and Elfassy, 2016, p. 59), to command respect and trust from followers. *Care* entails harboring an unselfish intent toward others so as facilitating their growth and development (Greenleaf, 1977; Shek and Li, 2015).

The Service Leadership Model builds on and complements other existing leadership paradigms such as servant leadership, ethical leadership, and transformational leadership (see Shek et al., 2015a for a thorough review). First, as discussed earlier, contrary to the servant leadership model deemphasizing one's own needs (Greenleaf, 1970; Russell and Stone, 2002), effective service leadership appreciates self-serving endeavors to develop one's capacity and eagerness to satisfy others' needs. Second, while the ethical leadership model emphasizes moral *Character* (Brown and Treviño, 2006), *Competence* (Shek et al., 2015a) and service provision on the "self" and "others" levels (Mendonca, 2001), how *Care* impacts leadership effectiveness remains under-addressed (Shek et al., 2015a). Third, transformational leaders motivate the pursuit of collective goals at the expense of personal interest, and in so doing these leaders help followers fulfill their potential through idealized influence, inspirational motivation, intellectual stimulation, and individualized considerations (Bass, 1990; Avolio et al., 1999). Transformational leadership theory has limited coverage on *Competence* and *Care* as the determinants of leadership success (Shek et al., 2015a).

In a nutshell, the Service Leadership Model incorporates several core features of related leadership paradigms and attempts to build up an integrative perspective in leadership (Shek et al., 2015a). Such a perspective inspires the education of a generation of new leaders that can thrive in this service era (Shek and Chung, 2015; Shek et al., 2015c, 2017).

Service Leadership Education in Hong Kong

As one of the most important outcomes of higher education, leadership of university students is highly regarded by both universities and employers (Bacon et al., 1979). However, a discrepancy exists between employers' expectation and what university students could demonstrate in service economies (Shek et al., 2017). Such a discrepancy results in a mismatch in recruitment, low job satisfaction and even mental burnout amongst the existing staff (Towers Watson, 2012). Thus, Po Chung, the co-founder of DHL International and the incumbent chairperson of the Hong Kong Institute of Service Leadership & Management Limited (HKI-SLAM), put forth the Service Leadership Model with a vision to nurture a generation of emergent service leaders who are not only competent, but are also moral and caring (Shek et al., 2017).

To promote quality leadership education conducive to students' personal growth and employability, Chung argued passionately for the need to incorporate formal training based on the Service Leadership Model into the curriculum of undergraduates in Hong Kong (Chung, 2015; Shek et al., 2015c). With the financial support of the Victor and William Fung Foundation and the collaborative effort from the HKI-SLAM and universities financed by the University Grants Committee (UGC), a multi-year project entitled "Fung Service Leadership Education Initiative (FSLEI)" was implemented in eight UGC-funded universities in Hong Kong. Based on the Service Leadership and Management (SLAM) curriculum framework proposed by the Hong Kong Institute of Service Leadership and Management Limited [HKI-SLAM] (2013), all institutions under the FSLEI independently developed programs and curriculum materials that facilitate learning of service leadership at the undergraduate level (Shek and Chung, 2015). While it is important to develop service leadership curriculum materials and training programs, it is equally important to develop objective measures of service leadership qualities (Shek and Chung, 2015). Unfortunately, the paucity of validated assessment tools on service leadership in the Chinese context (Shek et al., 2017) has hindered meaningful analyses on the effectiveness of service leadership education under the FSLEI (Shek and Lin, 2015b, 2017).

Against such a backdrop, the research team at a Hong Kong university initiated a multi-year project entitled 'Development and validation of measures based on the Service Leadership Model' (Shek et al., 2017). This project entailed the construction and validation of three scales, each of which constituted a parameter of success of an educational program (Shek and Lin, 2017) pertaining to one's *Attitude*, *Behavior*, and *Knowledge* on the Service Leadership Model (Shek et al., 2017). Some related publications can be seen elsewhere (e.g., Shek et al., 2018b,c,f; Shek and Chai, 2019). This paper primarily discusses the findings of a large-scale validation study on the Service Leadership Behavior Scale, which was designed to measure one's exhibited behavioral attributes characteristic of a service leader.

Service Leadership Behavior Scale

As part of the research program (Shek et al., 2017), the Long-Form Service Leadership Behavior Scale (SLB-LF-97) was developed primarily based on the SLAM curriculum framework (Hong Kong Institute of Service Leadership and Management Limited [HKI-SLAM], 2013), *25 Principles of Service Leadership* (Chung and Bell, 2015), *12 dimensions of a Service Leader* (Chung and Elfassy, 2016), and other published works from the leadership literature (e.g., Wielkiewicz, 2000; Ho and Nesbit, 2009). Initially, the SLB-LF-97 contained the following proposed domains: 3-Cs model (*Competence*, *Character* and *Care*), service provision, commitment to continuous improvement, and distributed leadership.

The SLB-LF-97 was administered in a preliminary validation study involving 231 university students (Shek et al., 2018b), where the results informed the retention of 65 items forming a short-form of the scale (SLB-SF-65). The SLB-SF-65 included 12 factors: problem-solving, self-leadership and life-long learning, non-cognitive intrapersonal competences, distributed leadership, integrity, care provision, concern, self-reflection, service provision, positive social relationship, communication skills, and fairness (Shek et al., 2018b). Both the SLB-LF-97 and the SLB-SF-65 exhibited excellent reliability (αs > 0.95) and robust convergent validity, with the latter evidenced by the significant and positive correlation with a host of theoretically relevant constructs such as servant leadership (r = 0.78) and leadership self-efficacy (r = 0.55) (Shek et al., 2018c). Nonetheless, the dimensionality of the SLB-SF-65 remained to be ascertained owing to the relatively modest sample size (N = 231). The background, conceptual model and steps involved in the development of different forms of Service Leadership Behavior Scales are outlined in Shek et al. (2018e).

Objectives of the Present Study

Utilizing the data from a validation study involving 4,486 undergraduates from eight UGC-funded universities, the present study sought to build upon the abovementioned preliminary validation study (Shek et al., 2018c) in two ways. First, following the commonly adopted two-step dimensionality analysis (Park, 2014; Besnoy et al., 2016) involving an exploratory factor analysis (EFA) followed by a confirmatory factor analysis (CFA), the present study attempted to examine the dimensionality of the SLB-SF-65. Second, via the utilization of a much larger sample alongside several well-validated external criterion measures adopted in the study of Shek et al. (2018c), the present study attempted to further establish the reliability and convergent validity of the SLB-SF-65. Based on Shek et al.'s (2018c) initial findings, this study constituted a pioneer effort to construct and validate an objective assessment tool on service leadership in a Chinese context. The present findings contribute to the scanty literature of service leadership evaluation in the Chinese context (Shek and Lin, 2015b, 2017) and serve to produce a valuable instrument to assess learning outcomes of service leadership training programs (Shek and Chung, 2015).

In the present study, evaluation of factorial validity of the SLB-SF-65 involved two steps, with the dataset (N = 4,486) randomly split into two halves (subsets A and B) to facilitate both the EFA and the CFA. The EFA performed on subset A (N = 2,246) resulted in a stable and valid initial six-factor, 48-item solution (SLB-SF-48, see **Figure 1**), which was consistent with the original conceptual model. Details pertaining to the EFA were reported in Shek et al. (2018c). The six factors, each of which formed a subscale on the basic dimensions of service leadership, were accordingly named (a) Self-improvement and Self-reflection (12 items), (b) People and Principles Orientation (12 items), (c) Resilience (8 items), (d) Social Competence (7 items), (e) Problem-Solving (6 items), and (f) Mentorship (3 items). In this paper, this six-factor solution was then subjected to a CFA performed on subset B (N = 2,240), with the objective to evaluate how this proposed model fit the rest of the data and stability of the factor structure.

MATERIALS AND METHODS

The data were derived from a research project on service leadership involving eight UGC-funded universities in Hong Kong. Students were invited to participate in the survey via an electronic platform. The data were collected between March and June, 2017. During the survey, the purpose of this study, the principles of voluntary participation and withdrawal, and the compensation arrangement were explained on the survey webpage and the invitation documents. Students were asked to indicate their acceptance or refusal to join the study on the opening page. We rewarded each participant a supermarket gift voucher valued at HK$100 (US$12.80).

Procedures

In total, 4,555 completed responses were retrieved. Three steps were performed for data cleaning. First, we removed six cases in which students declined to participate. Second, 30 cases were excluded because either they had completed the questionnaire designed for universities other than their own, or they revealed themselves as non-undergraduates in open-ended questions. Third, after reviewing respondents' student identity number (which is anonymous to the Research Team), 33 cases with multiple participation were removed from the sample. Ultimately, 4,486 cases were retained as the working sample.

Profiles of the Respondents

Among the 4,486 students, 1,517 were males and 2,969 were females. The majority of the sample were aged 20–24 years (68.4%; mean age = 20.47 years, SD = 1.67), had previous work experience (91.4%), and assumed the leadership position before (61.4%). Most participants had not received credit- or non-credit-bearing training in service leadership before (74.3 and 82.0%, respectively), and claimed to know "a little" or "some" about service leadership (75.0%).

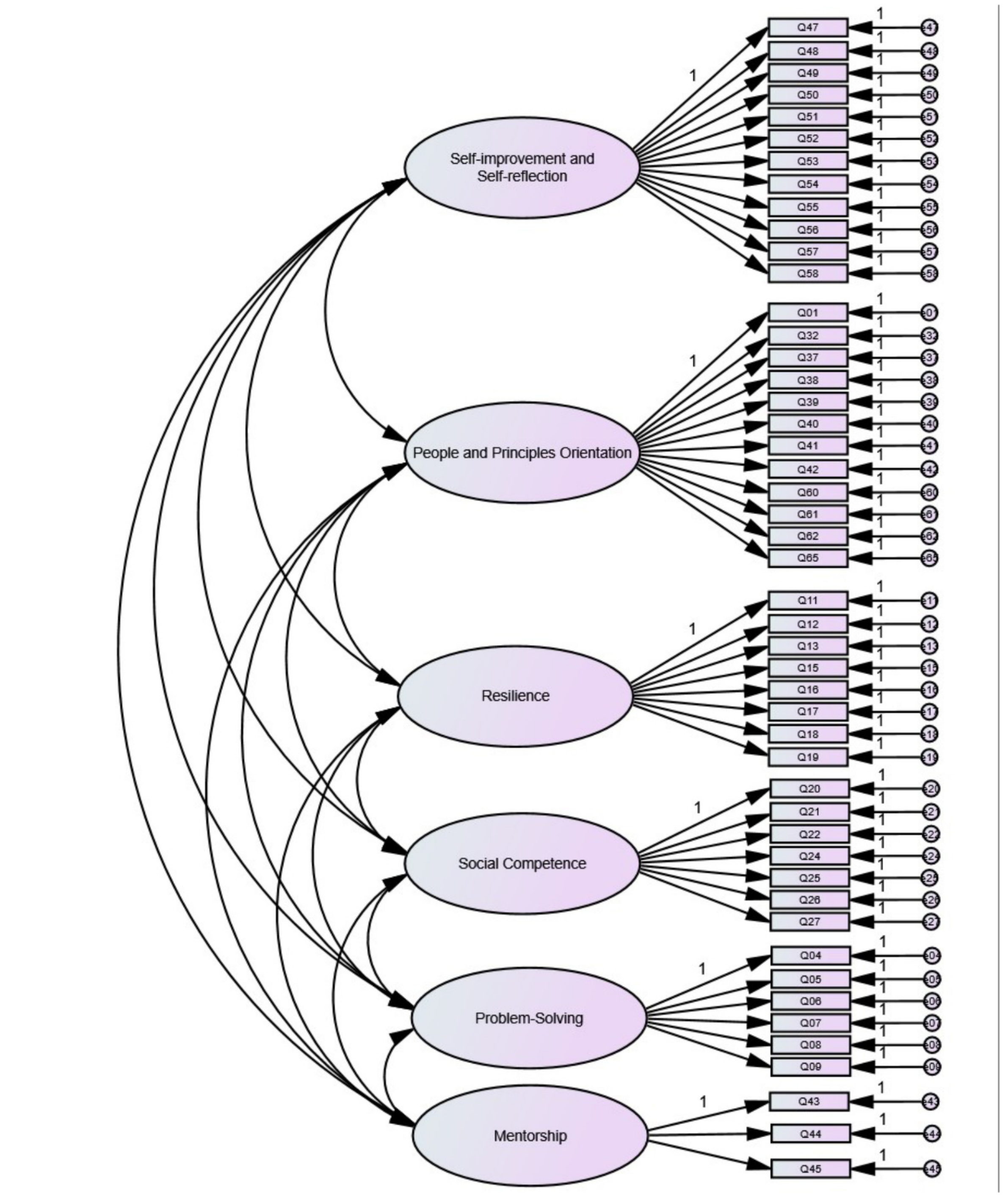


FIGURE 1 | The initial six-factor, 48-item factorial structure (Model 0; i.e., SLB-SF-48).

Instruments

Assessment of Service Leadership Qualities

The Long-Form Service Leadership Behavior Scale (SLB-LF-97) was designed to measure the behavioral attributes of an effective service leader (Shek et al., 2017). The 97 scale items were developed based on the general leadership literature (e.g., Wielkiewicz, 2000; Ho and Nesbit, 2009), publications based on the Service Leadership Model (e.g., Chung and Bell, 2015; Shek et al., 2015c; Chung and Elfassy, 2016) and the SLAM curriculum framework (Hong Kong Institute of Service Leadership and Management Limited [HKI-SLAM], 2013), with four domains, including the 3-Cs model (*Competence*, *Character* and *Care*),

TABLE 1 | Sample items of the Short-Form Service Leadership Behavior Scale (SLB-SF-65).

Items	Very Dissimilar to Me	Moderately Dissimilar to Me	Slightly Dissimilar to Me	Slightly Similar to Me	Moderately Similar to Me	Very Similar to Me
Sample item 1. I try to serve others without regard to their positions.	1	2	3	4	5	6
Sample item 2. I refuse to give in without a fight amidst adversity.	1	2	3	4	5	6
Sample item 3. I have no problem working with others.	1	2	3	4	5	6

All sample items were slightly re-phrased to avoid practice effect.

service provision, commitment to continuous improvement, and distributed leadership. The SLB-LF-97 was validated in a study involving 231 students from a university in Hong Kong (Shek et al., 2018b). The findings suggested the retention of 65 items to form the SLB-SF-65, which was employed in the present study. The dimensions derived are generally consistent with the original conceptual model. Each item of the SLB-SF-65 describes a specific leadership behavior where the respondents evaluate how well each item describes their leadership behavior (see **Table 1** for sample items). A six-point Likert scale was used (1 = very dissimilar; 6 = very similar). Both the SLB-LF-97 and the SLB-SF-65 recorded excellent internal consistency (αs > 0.95; mean inter-item correlations > 0.25) in the previous validation study (Shek et al., 2018c).

The research also entailed the construction of scales designed to assess individuals' knowledge of the Service Leadership Model (Shek et al., 2017, p. 167) as well as their attitudes and beliefs about desired leadership qualities (Shek et al., 2017, p. 212). In the present study, the shortened final versions of these two scales were administered.

Short-Form Service Leadership Knowledge Scale (SLK-SF-40)

The Service Leadership Knowledge Scale was developed based on the SLAM curriculum framework (Hong Kong Institute of Service Leadership and Management Limited [HKI-SLAM], 2013) and the literature on service leadership (e.g., Shek et al., 2015c; Chung and Elfassy, 2016). Participants' responses to the original 200 items were coded based on accuracies (1 = correct; 0 = incorrect). Based on a criterion-validation study involving 160 Hong Kong university students (Shek and Lin, 2017), 50 items were retained to form the shortened scale (SLK-SF-50). Then the SLK-SF-50 was administered in a large-scale validation study, of which the results suggested the removal of additional 10 items to form the final SLK-SF-40 (Shek et al., 2018d). **Table 2** illustrates several sample items of the final SLK-SF-40 administered in the present validation study.

Short-Form Service Leadership Attitude Scale (SLA-SF-46)

The Long-Form Service Leadership Attitude Scale was developed based on the Service Leadership Model (Shek et al., 2015b, 2018f) and the leadership literature (e.g., Page and Wong, 2000; Kopelman et al., 2008). Each of the original 132 statements presents a viewpoint on the nature of leadership and how a leader ought to conduct him/herself, where participants evaluated the extent to which they concurred with each item (Shek et al., 2017). A six-point Likert scale was used (1 = strongly disagree; 6 = strongly agree). Based on findings from an unpublished, quasi-experimental validation study involving 200 students from a university in Hong Kong, a shortened version of the survey containing 73 items was formed (SLA-SF-73). The SLA-SF-73 was further refined based on Exploratory Factor Analyses and Confirmatory factor analyses by using a large-scale sample (Ma et al., 2018; Shek and Chai, 2019). The final SLA-SF-46 used in the present study possesses excellent internal consistency (α = 0.93, mean inter-item correlations = 0.27). Sample items of the SLA-SF-46 are shown in **Table 3**.

The present study is primarily concerned with the validation findings for the SLB-SF-65. Details in relation to the validation of the SLA-SF-73 and the SLK-SF-50 are discussed in two separate papers (Shek et al., 2018d,f).

External Criterion Measures

Four external criterion scales adopted from the personality and leadership literature were used to gauge the convergent validity of the SLB-SF-65. These included the Revised Servant Leadership Profile (RSLP), Moral Self-Concept Scale (MSC), Leadership Efficacy Scale (LEF), and the Interpersonal Reactivity Index (IRI).

TABLE 2 | Sample items of the Short-Form Service Leadership Knowledge Scale (SLK-SF-40).

Items	Options	Correct answer
Sample item 1: A manager under the service economy wants to hire someone. Based on the Service Leadership Model, which of the following advice would you give him/her?	(A) Hire for qualifications, train for character (B) Hire for character, train for skills (C) Hire for attitude, train for character (D) Hire for efficiency, train for mindset	B
Sample item 2: Meg devoted herself to a career in relieving people of their hunger, isolation, and poverty. Which dimension of character strengths was shown by Meg's devotion?	(A) Justice (B) Courage (C) Humanity (D) Temperance	C

All sample items were slightly re-phrased to avoid practice effect.

TABLE 3 | Sample items of the Short-Form Service Leadership Attitude Scale (SLA-SF-46).

Items	Strongly Disagree	Disagree	Slightly Disagree	Slightly Agree	Agree	Strongly Agree
Sample item 1: Good leaders serve with a genuine heart.	1	2	3	4	5	6
Sample item 2: Good leaders give high priority to ethical issues.	1	2	3	4	5	6

All sample items were slightly re-phrased to avoid practice effect.

The RSLP was developed by Wong and Page (2003) to examine servant leadership. In this study, we selected five factors of the RSLP, which included 20 items that were highly relevant to the SLAM curriculum framework (Hong Kong Institute of Service Leadership and Management Limited [HKI-SLAM], 2013). These five factors are empowering and developing others (five items), serving others (seven items), open, participatory leadership (two items), inspiring leadership (two items), and courageous leadership (four items). The RSLP demonstrated excellent reliability in the present study (α = 0.94, mean inter-item correlations = 0.46).

The MSC was developed by Cheng (2005) to measure young people's self-appraisal on morality. The dimensions of MSC include conduct and virtues, self-control and disciplines, and altruism. All these aspects are crucial to how a service leader conducts himself/herself (Chung and Bell, 2015). The MSC presented good internal consistency in this study (α = 0.83, mean inter-item correlations = 0.44).

The LEF was developed by Murphy (1992) to examine one's level of confidence in his/her capacity to lead effectively. The LEF showed an acceptable internal consistency metrics (α = 0.70, mean inter-item correlations = 0.24).

The IRI was developed to assess empathy (Davis, 1983). In this study, we selected 14 items from two subscales of IRI, including empathic concern (IRI-EC, seven items) and perspective taking (IRI-PT, seven items). These two subscales are closely related to the qualities of an effective service leader (Chung and Elfassy, 2016). The IRI also showed good internal consistency in the present study (α = 0.74).

Analysis

Factorial Validity

Both exploratory (EFA) and confirmatory factor analysis (CFA) were involved in the validation study. While EFA provides preliminary evidence of a theoretical factorial solution (Shek et al., 2018c), CFA serves to verify the solution and validate the construct of the instrument (Besnoy et al., 2016). This two-step analytic approach has been commonly adopted to establish factorial validity of an instrument (e.g., Park, 2014; Wu and Mohi, 2015; Swami et al., 2017). SPSS version 24.0 (IBM) was utilized to administer the EFA and analyses of reliability and convergent validity. Mplus version 6.12 (Muthén and Muthén, 1998–2010) was used to perform the CFA.

As mentioned above, EFA was conducted on the SLB-SF-65 using a principal component analysis (PCA) with varimax rotation. Related findings suggested a six-factor structure of the trimmed scale (i.e., SLB-SF-48), which retained 48 items with factor loadings larger than 0.50. Besides, identical PCAs were performed on subsets A (N = 2,246) and B (N = 2,240). Tucker's coefficients of congruence (r_c) were used to evaluate the factor structure stability across the two subsets. SLB-SF-48 was revealed to be internally consistent and have a stable factorial structure. The item loadings of all 48 items ranged from 0.50 to 0.76. Details regarding the EFA and the steps involved in forming the initial 48-item behavior scale were reported in another paper (Shek et al., 2018c). The present paper primarily reports the findings of the CFA performed on the subset B (N = 2,240), internal consistency, convergent and factorial validity of the final version of the Service Leadership Behavior Scale (SLB-SF-38).

Before performing the main analyses, we conducted a preliminary screening to examine the skewness and kurtosis of the variables involved. Chou and Bentler's (1995) criteria was adopted (skewness < |2|; kurtosis < |7|). Then we administered the multigroup CFA (MGCFA) to establish measurement invariance of the final model. A series of MGCFAs were conducted following the steps suggested by van de Schoot et al. (2012), which specified configural, metric, scalar and error variance invariance models to be examined. The MGCFAs were performed on three pairs of subsamples under subset B (N = 2,240). One pair involved males (N = 728) versus females (N = 1,498), the second pair included "odd" (N = 1,120) versus "even" (N = 1,120) groups based on case number, and the third pair included "young" (N = 1,120) versus "old" (N = 1,120) groups based on student age. Due to length constraints and the similarity of the analyses between gender and age groups, the present study mainly reported the detailed information of measurement invariance tests on the first two pairs of subsamples.

The model fit was examined by indices including the chi-square (χ^2), comparative fit index (CFI), Bentler-Bonett Non-Normed fit index (NNFI), root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). We adopted the cutoff of 0.90 for both CFI and NNFI as indicators of adequate fit (Kline, 2005; Awang, 2012; van de Schoot et al., 2012). Regarding RMSEA and SRMR, a value below 0.80 and 0.10, respectively, should represent reasonable fit (Byrne, 1998; Hirsh, 2010). Considering that χ^2 test is sensitive to sample size and model complexity, we adopted difference-in-CFI (ΔCFI) as the main invariance test indicator (Cheung and Rensvold, 2002). Particularly, as proposed by Cheung and Rensvold's (2002), a ΔCFI below (or equal to) 0.01 suggests invariance (Schmitt and Kuljanin, 2008; Byrne, 2010). Additionally, modification indices (M.I.s) of items were reviewed upon marginal model fit. Some researchers suggested that items with extreme M.I.s (i.e., >40.0) should be dropped (Anderson and Gerbing, 1988, p. 417).

Reliability and Convergent Validity

Cronbach's alpha values and mean inter-item correlations were used as the indicators of reliability of the behavior scale and the subscales derived. We also examined the convergent validity of the behavior scale in terms of its correlation with relevant constructs such as servant leadership and empathy measured by external measures (e.g., RSLP, IRI). Specifically, considering that servant leadership, moral self-concept, leadership efficacy and empathy were all key behavioral prerequisites of a service leader (see Chung and Bell, 2015; Chung and Elfassy, 2016), we hypothosized a positive and significant correlation between the service leadership behavior scale and the RSLP (Hypothesis 1), MSC (Hypothesis 2), LEF (Hypothesis 3), and IRI (Hypothesis 4), respectively.

The convergent validity of the behavior scale could be further evidenced by its correlation with the SLA-SF-46 and the SLK-SF-40. Since all three scales were constructed to examine different facets of service leadership, we predicted a positive and significant correlation between the behavior scale (and its subscales) with both the SLA-SF-46 (Hypothesis 5) and the SLK-SF-40 (Hypothesis 6).

RESULTS

Data Screening and Descriptive Statistics

As detailed in **Table 4**, Cronbach's alpha values and mean inter-item correlations showed good internal consistency of the initial six-factor solution (see **Figure 1**). No abnormal findings were found regarding each variable's means, standard deviation, univariate skewness and kurtosis values. In short, the descriptive analyses informed the normality of data distribution, rendering the use of Maximum Likelihood (ML) estimation method appropriate. The sample size of the present study (N = 2,240) was also adequately powered (MacCallum et al., 1999).

Factorial Validity Assessment

Factor Structure of the Initial Model: SLB-SF-48

Based on the original EFA solution, the findings revealed that the initial model (SLB-SF-48) fit the data reasonably well (RMSEA = 0.061; SRMR = 0.046), although some indices (CFI = 0.86; NNFI = 0.86) fell short of the recommended levels (Aquino and Reed, 2002). After reviewing the modification indices (M.I.s), we further removed 10 items reflecting double factor loadings or a strong residual covariance with other items or factors (see **Table 5**) (Anderson and Gerbing, 1988; Awang, 2012). The alpha values remained high when an item was removed from the scale (ranged from 0.853 to 0.925, see **Table 5**). The resultant six-factor, 38-item model (Model 1) was subjected to the second CFA.

Factor Structure of the Modified Model: SLB-SF-38

As detailed in **Table 6**, the fit indices considerably improved after the deletion of problematic items (CFI = 0.902; NNFI = 0.894; RMSEA = 0.056; SRMR = 0.045). The M.I.s of this 38-item model (Model 1) were further scrutinized. Three pairs of parameters indicated high covariance, including items Q04 and Q05 (M.I. = 239.75), Q18 and Q19 (M.I. = 150.34), and Q49 and Q50 (M.I. = 399.57).

Byrne (1998) contended that these extreme M.Is. may be attributed to the unique characteristics that these items shared in content. Accordingly, these three pairs of scale items were revisited. First, both items Q04 and Q05 refer to problem-solving. Second, items Q18 and Q19 measure specifically participants' adaptive coping strategies amidst adversity. Third, both items Q49 and Q50 tap into participants' mindset or competence in goal-setting. In a nutshell, all these observations pointed toward an overlap in content amongst the three pairs of items, which justified the inclusion of error correlations amongst these pairs (Shek and Yu, 2014). Consequently, three modified models were re-specified based on Model 1. More specifically, Model 2 included a correlation between errors of items Q04 and Q05; Model 3 built on Model 2 by incorporating an error covariance of items Q18 and Q19; Model 4 further added to Model 3 by co-varying the errors of items Q49 and Q50. **Table 6** presents the goodness-of-fit statistics of Model 1 to Model 4 so as the initial six-factor 48-item solution (Model 0).

All indices represented the adequate fit of Model 4 to the data (χ^2(647) = 4,496.31; CFI = 0.919; NNFI = 0.912, RMSEA = 0.052 [90% CI: 0.050–0.053]; SRMR = 0.046). The results of Chi-square tests showed that Model 2, Model 3 and Model 4 demonstrated significant improvement compared to Model 1, Model 2 and Model 3, respectively. We also referred to the difference-in-CFI (ΔCFI) indicator with reference to Cheung and Rensvold's (2002) proposed cutoff of $|0.01|$ as the benchmark. The results showed that Model 4 significantly improved than Model 1. As a result, Model 4 was accepted as the final model (SLB-SF-38, see **Figure 2**).

As shown in **Table 7**, the standardized factor loadings of all 38 items were above 0.50 ($p < 0.001$, two-tailed), and squared multiple correlations were greater than 0.25 ($p < 0.001$, two-tailed).

Invariance Tests Across Genders

Model 4 was tested separately by gender in Model 5 and Model 6 to gauge its factorial stability (Byrne, 1998; Shek and Ma, 2010). As shown in **Table 6**, both models demonstrated adequate fit to the data in both the male (Model 5: χ^2 (647) = 1,896.30; CFI = 0.922; NNFI = 0.915, RMSEA = 0.051 [90% CI: 0.048 to 0.054]; SRMR = 0.043) and female subsamples (Model 6: χ^2 (647) = 3,606.68; CFI = 0.906; NNFI = 0.898, RMSEA = 0.055 [90% CI: 0.054 to 0.057]; SRMR = 0.051). As illustrated in **Table 8**, all factor loadings and the squared multiple correlations in the two models were significant at $p < 0.001$, two-tailed.

As abovementioned, the invariance models were tested by the configural invariance model (Model 9), the metric invariance model (Model 10), the scalar invariance model (Model 11), and the error variance invariance model (Model 12). **Table 9** showed the results of the Chi-square tests, which revealed no significant difference between Model 9 and 10 ($\Delta\chi^2$ = 39.03, Δdf = 32, $p > 0.05$), but significant differences between Model 10 and 11 ($\Delta\chi^2$ = 187.30, Δdf = 38, $p < 0.001$), and between Model 11 and 12 ($\Delta\chi^2$ = 499.81, Δdf = 41, $p < 0.001$). As mentioned

TABLE 4 | Descriptive statistics and Reliability indices of the original 48-item model (SLB-SF-48).

	Factors						
	Reliability statistics		Descriptive statistics			Skewness	Kurtosis
Factors	**α**	**Mean inter-item correlations**	**Items**	**Mean**	***SD***	**Value**	**Value**
(1) Self-improvement and Self-Reflections (12 items)	0.930	0.527	Q47	4.56	1.00	−0.729	0.813
			Q48	4.65	0.94	−0.719	0.995
			Q49	4.50	0.98	−0.695	0.863
			Q50	4.42	1.06	−0.690	0.504
			Q51	4.72	0.88	−0.742	1.393
			Q52	4.66	0.95	−0.703	0.892
			Q53	4.61	0.93	−0.626	0.762
			Q54	4.56	0.98	−0.756	0.959
			Q55	4.67	0.97	−0.689	0.649
			Q56	4.70	0.94	−0.792	1.161
			Q57	4.61	0.95	−0.705	0.969
			Q58	4.60	0.98	−0.586	0.524
(2) People and Principles Orientation (12 items)	0.905	0.446	Q01	4.38	0.96	−0.780	1.135
			Q32	4.55	0.95	−0.659	0.842
			Q37	4.71	0.92	−0.903	1.736
			Q38	4.79	0.90	−0.800	1.163
			Q39	4.77	0.93	−0.756	0.974
			Q40	4.48	0.99	−0.680	0.685
			Q41	4.67	0.88	−0.706	1.153
			Q42	4.65	0.88	−0.640	0.982
			Q60	4.75	0.91	−0.887	1.635
			Q61	4.58	0.93	−0.849	1.479
			Q62	4.64	0.85	−0.638	1.117
			Q65	4.82	0.86	−0.807	1.406
(3) Resilience (8 items)	0.888	0.502	Q11	4.19	1.10	−0.513	−0.064
			Q12	4.19	1.09	−0.452	−0.089
			Q13	4.31	1.04	−0.587	0.331
			Q15	4.44	0.96	−0.624	0.740
			Q16	4.50	0.96	−0.526	0.468
			Q17	4.43	0.97	−0.530	0.480
			Q18	4.27	1.09	−0.586	0.226
			Q19	4.23	1.11	−0.514	0.025
(4) Social Competence (7 items)	0.898	0.556	Q20	4.65	0.92	−0.738	1.121
			Q21	4.64	0.94	−0.896	1.347
			Q22	4.70	0.91	−0.795	1.344
			Q24	4.51	0.93	−0.751	1.138
			Q25	4.36	1.00	−0.576	0.420
			Q26	4.39	0.96	−0.585	0.436
			Q27	4.48	0.97	−0.622	0.665
(5) Problem-Solving (6 items)	0.875	0.539	Q04	4.43	0.93	−0.463	0.305
			Q05	4.22	1.00	−0.467	0.191
			Q06	4.56	0.97	−0.614	0.585
			Q07	4.53	0.95	−0.544	0.487
			Q08	4.48	0.98	−0.629	0.565
			Q09	4.38	0.95	−0.610	0.726
(6) Mentorship (3 items)	0.847	0.647	Q43	4.35	0.98	−0.572	0.472
			Q44	4.19	1.04	−0.496	0.192
			Q45	4.18	1.07	−0.540	0.146
Service Leadership Behavior Scale (48 items; SLB-SF-48)	0.966	0.377	–	4.51	0.60	–	–

N = 2,240. α, Cronbach's alpha coefficients. SD, standard deviation.

TABLE 5 | Items removed from SLB-SF-48 due to extreme modification indices.

Factors	Items removed	α if an item is deleted	Modification indices (M.I.s) with items within the same factor	
			Items	Modification indices
Problem-solving	Q09	0.853	Q06	42.28
			Q08	193.93
Resilience	Q11	0.880	Q12	509.64
			Q15	76.25
			Q16	45.13
			Q17	46.10
Social Competence	Q25	0.880	Q21	67.80
			Q22	101.95
			Q24	89.73
			Q26	262.31
			Q27	80.86
Social Competence	Q26	0.885	Q20	99.04
			Q21	107.36
			Q22	88.42
			Q25	262.31
			Q27	176.58
People and Principles Orientation	Q39	0.896	Q38	156.48
			Q40	51.36
People and Principles Orientation	Q41	0.895	Q42	380.63
			Q60	45.39
People and Principles Orientation	Q61	0.898	Q42	42.86
			Q60	131.43
			Q62	194.29
			Q65	50.34
Self-improvement and Self-reflection	Q47	0.925	Q48	81.84
			Q50	180.82
			Q52	69.20
			Q54	60.44
Self-improvement and Self-reflection	Q53	0.924	Q49	47.03
			Q52	227.19
			Q54	227.11
			Q57	43.85
Self-improvement and Self-reflection	Q57	0.924	Q53	43.85
			Q55	52.08
			Q56	78.19
			Q58	179.73

Only M.I.s (with items within the same factor) larger than 40.00 were shown.

earlier, we followed Cheung and Rensvold's suggestion (Cheung and Rensvold, 2002) and referred to the value of Δ CFI.

As shown in **Table 9**, Model 9 in which no quality constraint was postulated fit adequately with the data (χ^2 (1,294) = 5,502.998; CFI = 0.912; NNFI = 0.904, RMSEA = 0.054 [90% CI: 0.052 to 0.055]; SRMR = 0.049), suggesting invariance of the overall factorial structure across genders. In Model 10, factor loadings were constrained to be equal across genders. The value of ΔCFI (<0.001) compared to Model 9 was below Cheung and Rensvold's (2002) proposed cutoff (0.01), suggesting invariance in factor loadings as well across genders.

In Model 11, equality constraints were placed upon both factor loadings and measurement intercepts across the male and female groups. The value of ΔCFI (0.004) denoted invariance in measurement intercepts of each item across genders (see **Table 9**).

Lastly, in Model 12 we constrained the error variance, factor loading, and measurement intercept of each variable to be equal across genders to establish error variance invariance model (Model 12). The value of ΔCFI (0.009, see **Table 9**) was again below 0.01, suggesting that same level of measurement error was present for each item between males and females (Milfont and Fischer, 2010, p. 115).

Invariance Tests Across Other Subsamples

Following Shek and colleagues' procedure (Shek and Ma, 2010, 2014; Shek and Yu, 2014), subset B (N = 2,240) was further divided into group "odd" (N = 1,120) and group "even" (N = 1,120) based on case number. Both groups were subjected to the identical set of invariance tests as reported above. As shown in **Table 6**, Model 4 fitted reasonably well with the dataset in both the odd (Model 7: χ^2 (647) = 2,683.30; CFI = 0.917; NNFI = 0.910, RMSEA = 0.053 [90% CI: 0.051 to 0.055]; SRMR = 0.047) and even groups (Model 8: χ^2 (647) = 2,837.68; CFI = 0.906; NNFI = 0.898, RMSEA = 0.055 [90% CI: 0.053–0.057]; SRMR = 0.050). These findings provided basis for the ensuing series of MGCFAs, which served to establish measurement invariance across the two subsamples.

In Model 13, no equality constraints were imposed. As illustrated in **Table 9**, the goodness-of-fit indices of Model 13 exhibited acceptable fit to the data (χ^2(1,294) = 5,520.98; CFI = 0.912; NNFI = 0.904, RMSEA = 0.054 [90% CI: 0.053 to 0.055]; SRMR = 0.048), suggesting configural invariance. We further constrained the factor loadings to be equal in Model 14 and compared it with the baseline Model 13. The result of χ^2 test was significant at the 0.05 level ($\Delta\chi^2 = 46.66$, $\Delta df = 32$, $p < 0.05$). The resultant value of ΔCFI (<0.001) provided support for the metric invariance across the two subsamples. In Model 15, equality constraints were further placed on the measurement intercepts of all items. The χ^2 test showed a non-significant result ($\Delta\chi^2 = 40.56$, $\Delta df = 38$, $p > 0.05$). Likewise, the value of ΔCFI (<0.001) derived from the comparison between Model 14 and Model 15 conveyed scalar invariance. In Model 16 the error variance, factor loading and measurement intercept were held equal for every item across both subsamples. Although the χ^2 test showed a significant difference between Model 16 and 15 ($\Delta\chi^2 = 76.56$, $\Delta df = 41$, $p < 0.001$), the resultant value of ΔCFI (0.002) remained trivial by Cheung and Rensvold's (2002) standard, signaling error variance invariance of the final factorial solution (SLB-SF-38) as displayed in **Figure 2**.

Besides, we also examined the measurement invariance across age groups by dividing subset B (N = 2,240) into two groups based on student age. The "Young" Group (N = 1,120, mean age = 19.17 years, SD = 0.76) and "Old" Group (N = 1,120, mean

TABLE 6 | Goodness-of-fit statistics for the modified CFA models.

Model	Modifications	Comparative models	χ^2	$\Delta\chi^2$	*df*	Δdf	*p*	CFI	NNFI	RMSEA (90% CI)	SRMR
0	Original model (Six-factor SLB-SF-48)		9,939.86		1,065			0.864	0.856	0.061 (0.060 – 0.062)	0.046
1	10 items deleted (see **Table 5**) from Model 0		5,297.81		650			0.902	0.894	0.056 (0.055 – 0.058)	0.045
		1 versus 0		4642.05		415	<0.001				
2	Model 1 + correlated errors of Q04 and Q05		5,062.22		649			0.907	0.900	0.055 (0.054 – 0.057)	0.046
		2 versus 1		235.59		1	<0.001				
3	Model 2 + correlated errors of Q18 and Q19		4,912.93		648			0.910	0.903	0.054 (0.053 - 0.056)	0.046
		3 versus 2		149.29		1	<0.001				
4	Model 3 + correlated errors of Q49 and Q50		4,496.31		647			0.919	0.912	0.052 (0.050 – 0.053)	0.046
		4 versus 3		416.62		1	<0.001				
5	Model 4: Males (*N* = 742; separate testing)		1,896.30		647			0.922	0.915	0.051 (0.048 – 0.054)	0.043
6	Model 4: Females (*N* = 1,498; separate testing)		3,606.68		647			0.906	0.898	0.055 (0.054 – 0.057)	0.051
7	Model 4: Odd (*N* = 1,120; separate testing)		2,683.30		647			0.917	0.910	0.053 (0.051 – 0.055)	0.047
8	Model 4: Even (*N* = 1,120; separate testing)		2,837.68		647			0.906	0.898	0.055 (0.053 – 0.057)	0.050
Criterion for goodness-of-fit			–		–			≥0.90	≥0.90	<0.08	<0.10

N_{whole} = 2,240. All χ^2 values were statistically significant at $p < 0.001$ (two-tailed); $\Delta\chi^2$, change in χ^2 compared to the previous model; Δdf, change in degrees of freedom compared to the previous model; CFI, Comparative Fit Index; RMSEA, Root Mean Square Error of Approximation; CI, Confidence Interval; NNFI, Bentler–Bonett Non-Normed Fit Index; SRMR, Standardized Root Mean Square Residual.

age = 21.71, *SD* = 1.24) were subjected to the same invariance tests mentioned above. Same as gender invariance, the resultant values of ΔCFI ($\leq$0.01) also supported configural, metric, scarlar and error variance invariance of the factorial structure between the two age groups.

In summary, the present findings provided strong support for the factorial validity of the 38-item Service Leadership Behavior Scale (SLB-SF-38). Apart from exhibiting adequate fit to the data, the strong factorial stability of the SLB-SF-38 was underscored by the series of invariance test performed based on groups defined by gender and age as well as with randomly assigned subjects. Specifically, measurement invariance of the SLB-SF-38 was supported in terms of configural, metric, scalar, and error variance invariance.

Reliability of the Measures

As indicated in **Table 10**, the SLB-SF-38 showed excellent reliability (α = 0.96, mean inter-item correlations = 0.38). All its six subscales also demonstrated good to excellent reliability in the present study (αs > 0.84, mean inter-item correlations > 0.35). The inter-correlations among the SLB-SF-38 and the subscales ranged from 0.42 to 0.87 ($p < 0.001$, two-tailed). These findings underscored the strong internal consistency of the SLB-SF-38 and the subscales.

Convergent Validity Assessment

Correlation With External Criterion Measures

As shown in **Table 11**, consistent with Hypotheses 1 to 4, correlational findings revealed the significant ($p < 0.001$, two-tailed) and positive association between the SLB-SF-38 (inclusive of all subscales) and the RSLP (*r*s ranging from 0.49 to 0.79), MSC (*r*s ranging from 0.37 to 0.66), LEF (*r*s ranging from 0.37 to 0.52) and IRI (*r*s ranging from 0.20 to 0.55). These findings provided convergent evidence for the validity of the SLB-SF-38, given that this scale was moderately related to several constructs outlining the behavioral characteristics of a service leader (Chung and Elfassy, 2016).

Correlation With Other Service Leadership Measures

Furthermore, findings of correlational analyses between the SLB-SF-38 and the final versions of the Service Leadership Attitude (SLA-SF-46) and Knowledge (SLK-SF-40) Scales are summarized in **Table 12**. Discussions in relation to the validation of the eight-factor SLA-SF-46 as well as the one-factor SLK-SF-40 are featured in two other papers. The SLB-SF-38 was overall moderately and positively linked to the SLA-SF-46 (*r* = 0.58) and also positively linked to the SLK-SF-40 (*r* = 0.19). The subscales of the SLB-SF-38 were also correlated positively and significantly with both the SLA-SF-46 and the SLK-SF-40. Although some occasional non-significant and unexpected results were observed, the results of correlational analyses supported Hypotheses 5 and 6.

To conclude, the present findings offered solid and consistent evidence for the construct validity of the SLB-SF-38. The main scale and the six subscales were correlated with a series of well-validated measures developed to examine constructs related to service leadership. Besides, the SLB-SF-38 and the subscales were also correlated with Service Leadership Attitude Scale and Service Leadership Knowledge Scale, which assessed the different

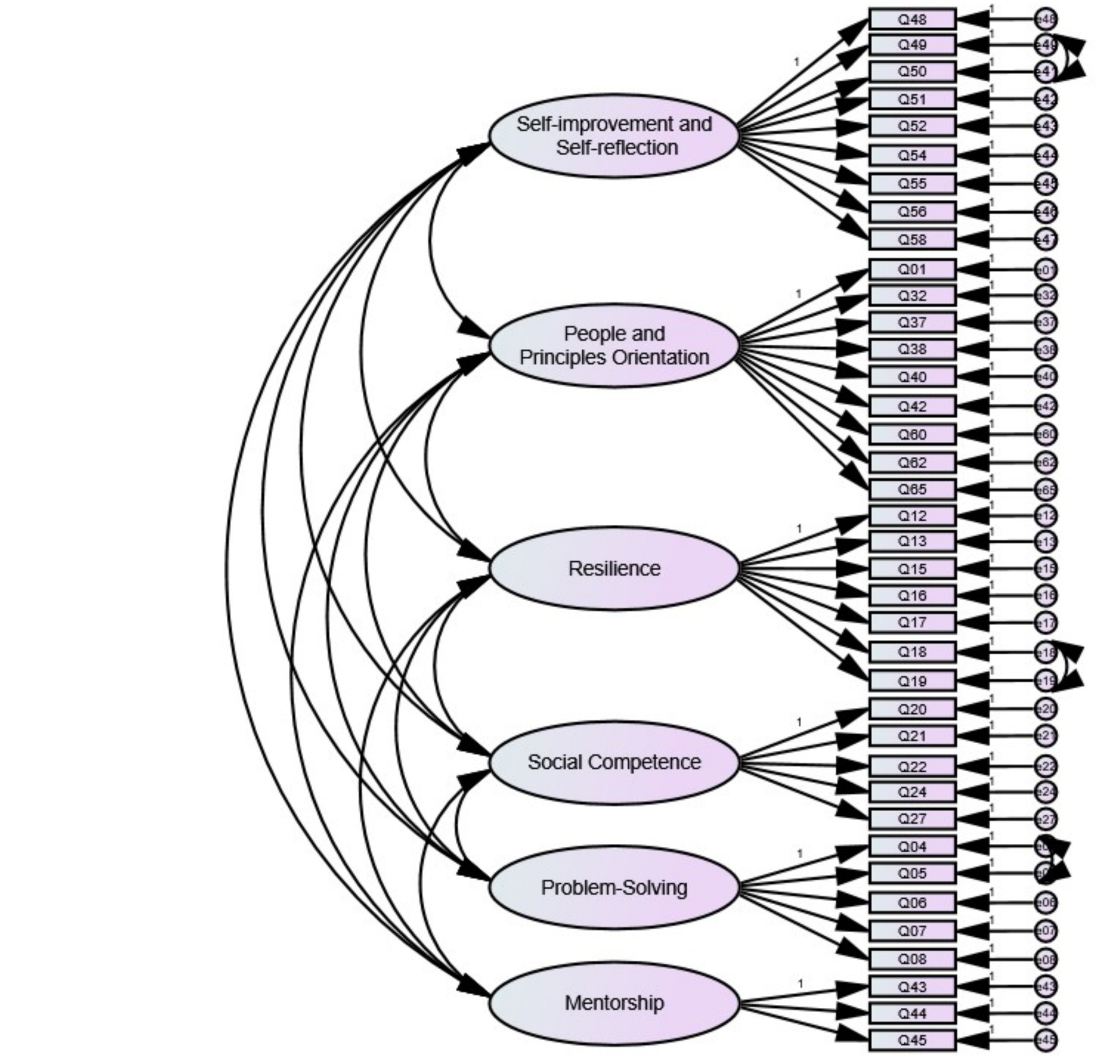


FIGURE 2 | The final six-factor factorial model (Model 4; i.e., SLB-SF-38) of the Service Leadership Behavior Scale.

dimensions of the same underlying construct. Thus, the SLB-SF-38 is shown to be a valid and reliable measurement tool of the behavioral characteristics of a service leader.

DISCUSSION

The present study attempted to examine the reliability, convergent validity and dimensionality of the Short-Form Service Leadership Behavior Scale (SLB-SF-65) based on a large sample of Hong Kong undergraduates. The findings suggested the retention of 38 items, which can be grouped under six dimensions including "Self-improvement and Self-reflection," "People and Principles Orientation," "Resilience," "Social Competence," "Problem-Solving," and "Mentorship." The results of multi-group CFA supported the stability of this factorial structure. Both the SLB-SF-38 and the six subscales presented good internal consistency and robust convergent validity. In short, this study validated the SLB-SF-38 as a sound assessment tool to evaluate the behavioral attributes of service leaders.

There are several strengths of the present study. First, the development of the scales were driven by the Service Leadership Model, which has been extensively covered in the literature and shown to be beneficial to university students in Hong Kong (Shek and Chung, 2015; Shek et al., 2017). Second, the present study employed a large sample which accounted for 5.36% of the total 84,388 Hong Kong undergraduates in the 2016/17 academic year (University Grants Committee [UGC], 2017). This large sample contributed to the robust findings (Biau et al., 2008). Third, the present study constructed an objective and psychometrically sound measurement tool to the leadership and youth development literature. Fourth, this study validated an objective measurement assessing service leadership behaviors in a Chinese context with an important role in the global service economy.

TABLE 7 | Standardized factor loadings for the six subscales of SLB-SF-38 (Model 4).

Subscales	Items	Factor loadings	SMC
(1) Self-improvement and Self-reflection	Q48	0.72	0.52
	Q49	0.65	0.42
	Q50	0.62	0.39
	Q51	0.76	0.58
	Q52	0.75	0.57
	Q54	0.74	0.54
	Q55	0.76	0.58
	Q56	0.80	0.65
	Q58	0.71	0.51
(2) People and Principles Orientation	Q01	0.51	0.26
	Q32	0.62	0.38
	Q37	0.71	0.51
	Q38	0.73	0.53
	Q40	0.63	0.40
	Q42	0.70	0.49
	Q60	0.69	0.47
	Q62	0.63	0.39
	Q65	0.65	0.43
(3) Resilience	Q12	0.62	0.38
	Q13	0.71	0.50
	Q15	0.77	0.59
	Q16	0.78	0.60
	Q17	0.78	0.61
	Q18	0.67	0.45
	Q19	0.66	0.44
(4) Social Competence	Q20	0.79	0.63
	Q21	0.83	0.69
	Q22	0.81	0.66
	Q24	0.72	0.51
	Q27	0.64	0.41
(5) Problem-Solving	Q04	0.65	0.42
	Q05	0.64	0.41
	Q06	0.79	0.62
	Q07	0.79	0.63
	Q08	0.74	0.55
(6) Mentorship	Q43	0.75	0.56
	Q44	0.86	0.74
	Q45	0.82	0.67

N = 2,240. SMC, Squared multiple correlations. All standardized factor loadings (STDYX metrics) and SMC were statistically significant at $p < 0.001$ (two-tailed).

The present six dimensions aligned well with the Service Leadership Model. First, the factor "Self-improvement and Self-reflection" (nine items) emphasizes the importance of reviewing and improving one's own leadership behavior as a continuous quest (Chung and Bell, 2015, p. 59). The second factor "People and Principles Orientation" (9 items) is concerned with having a set of personal code of ethics and treating others with care (Chung and Elfassy, 2016). This dimension is consistent with the morality, trust, fairness and respect emphasized in Service Leadership Model. Third, the dimension "Resilience" (seven items) measures an individual's ability to effectively respond toward stress, difficulty, and other unpleasant events in life (Shek and Lin, 2015c). This dimension can be conceptualized as an intrapersonal competence that enhances leadership effectiveness (Patel, 2012; Hatler and Sturgeon, 2013). Therefore, resilience constitutes an essential behavioral attribute of an effective service leader, and it is definitely a key component of service leadership education (Shek and Leung, 2015). The fourth factor "Social Competence" (five items) covers three aspects on one's capacity to effectively handle social interactions. These aspects include the ability to get along with other people, to build and accordingly maintain close relationships, and to behave appropriately in social settings (see Orpinas, 2010). This factor echoes the interpersonal competence outlined in Service Leadership Model. Fifth, the dimension "Problem-Solving" (five items) measures people's critical thinking when tackling difficult or complex issues (Altun, 2003). Problem-Solving falls into the category of intrapersonal competence as part of the service leadership education curriculum (Shek and Leung, 2015). Effective problem-solving is vital to leadership success (Mumford et al., 2000), and closely related to other intrapersonal competence such as emotion management (Mehrdad et al., 2011). Furthermore, service leaders may need to solve potentially conflicting needs of self, others, and the systems without compromising on morality. In this situation, critical thinking will help service leaders to see bigger picture and handle the problem in a timely manner (Jasovsky and Kamienski, 2007). Thus, the factor "Problem-Solving" underlies a dimension of behavioral attributes of service leadership. Lastly, the subscale "Mentorship" (three items) measures participants' capability and willingness to support other's development (Shek and Lin, 2015d), echoing the *Competence* and *Care* components highlighted in the Service Leadership Model. In short, the findings provide support for the "3-Cs" (*Competence*, *Character* and *Care*) of the Service Leadership Model. The results also echo the belief that both "being" (i.e., *Character* and *Care*) and "doing" (i.e., *Competence*) are important for effective leadership. The findings are pioneering in terms of constructing a validated measures of service leadership in Chinese societies.

The present study provides support for the developed tool on service leadership behavior. The findings enable cross-institutional analyses on curriculum effectiveness, and also offer robust empirical support for the Service Leadership Model (Shek and Chung, 2015; Shek et al., 2017). Theoretically speaking, the finings underscore the importance of the different dimensions of the measure as components of service leadership. This contributes to the development of the theory of service leadership.

The present study has several practical implications. First, the SLB-SF-38 can be employed to assess the impact of a service leadership training program. As students are expected to demonstrate an improvement in behavioral attributes of service leadership after completing the program, educators can use this tool to assess the change. Second, the dimensionality of the SLB-SF-38 can be used to refine service leadership education curriculum. Specifically, the curriculum materials for future service leadership training may be tuned to focus on the six dimensions identified. Third, the SLB-SF-38 can be used by

TABLE 8 | Complete standardized factor loadings and squared multiple correlations for Model 5 to Model 8.

	Model 5 (Males; N = 742)		**Model 6 (Males; N = 1,498)**		**Model 7 (Odd; N = 1,120)**		**Model 8 (Even; N = 1,120)**	
	Factor loadings	**SMC**	**Factor loadings**	**SMC**	**Factor loadings**	**SMC**	**Factor loadings**	**SMC**
Factor 1. Self-improvement and Self-reflection								
Q48	0.71	0.51	0.73	0.53	0.72	0.52	0.73	0.53
Q49	0.64	0.41	0.65	0.42	0.63	0.40	0.66	0.44
Q50	0.63	0.39	0.63	0.39	0.63	0.39	0.62	0.39
Q51	0.75	0.56	0.76	0.58	0.74	0.55	0.78	0.60
Q52	0.75	0.58	0.75	0.56	0.75	0.56	0.76	0.58
Q54	0.74	0.54	0.74	0.54	0.75	0.56	0.72	0.52
Q55	0.74	0.54	0.77	0.60	0.76	0.58	0.76	0.57
Q56	0.78	0.61	0.82	0.67	0.82	0.67	0.79	0.62
Q58	0.72	0.51	0.71	0.50	0.71	0.51	0.72	0.51
Factor 2. People and Principles Orientation								
Q01	0.59	0.35	0.46	0.21	0.52	0.27	0.51	0.26
Q32	0.65	0.42	0.59	0.34	0.64	0.41	0.59	0.35
Q37	0.74	0.55	0.69	0.48	0.73	0.53	0.71	0.50
Q38	0.75	0.57	0.71	0.50	0.74	0.55	0.72	0.52
Q40	0.68	0.46	0.60	0.36	0.64	0.41	0.63	0.39
Q42	0.73	0.54	0.67	0.45	0.70	0.49	0.70	0.48
Q60	0.70	0.58	0.68	0.46	0.69	0.48	0.68	0.47
Q62	0.64	0.67	0.62	0.38	0.63	0.40	0.62	0.39
Q65	0.65	0.65	0.65	0.43	0.66	0.44	0.65	0.42
Factor 3. Resilience								
Q12	0.58	0.34	0.64	0.41	0.60	0.36	0.63	0.40
Q13	0.68	0.46	0.73	0.53	0.72	0.52	0.69	0.48
Q15	0.77	0.60	0.76	0.58	0.76	0.57	0.78	0.61
Q16	0.75	0.57	0.79	0.63	0.77	0.59	0.79	0.62
Q17	0.78	0.61	0.78	0.60	0.81	0.65	0.75	0.56
Q18	0.69	0.47	0.66	0.43	0.67	0.45	0.67	0.45
Q19	0.66	0.44	0.66	0.44	0.71	0.50	0.61	0.37
Factor 4. Social Competence								
Q20	0.78	0.61	0.80	0.64	0.82	0.68	0.76	0.58
Q21	0.81	0.66	0.84	0.71	0.83	0.70	0.83	0.69
Q22	0.80	0.63	0.81	0.66	0.81	0.66	0.81	0.65
Q24	0.72	0.52	0.71	0.50	0.74	0.54	0.69	0.48
Q27	0.68	0.46	0.61	0.37	0.67	0.45	0.61	0.37
Factor 5. Problem-Solving								
Q04	0.65	0.42	0.65	0.43	0.66	0.44	0.64	0.41
Q05	0.67	0.45	0.63	0.39	0.64	0.41	0.65	0.42
Q06	0.78	0.61	0.79	0.63	0.76	0.58	0.82	0.66
Q07	0.78	0.62	0.80	0.64	0.77	0.59	0.81	0.66
Q08	0.73	0.54	0.75	0.56	0.75	0.56	0.74	0.55
Factor 6. Mentorship								
Q43	0.76	0.58	0.74	0.55	0.76	0.57	0.74	0.55
Q44	0.82	0.67	0.89	0.79	0.85	0.73	0.87	0.76
Q45	0.80	0.65	0.82	0.68	0.81	0.66	0.82	0.67

N_{whole} = 2,240. SMC, Squared multiple correlations. All standardized factor loadings (STDYX metrics) and SMC were statistically significant at $p < 0.001$ (two-tailed).

TABLE 9 | Summary of goodness-of-fit for invariance tests: multigroup comparisons.

Model description	Comparative model	χ^2	$\Delta\chi^2$	*df*	Δ*df*	Statistical significance	CFI	ΔCFI	NNFI	ΔCFI ≤ |0.01|?	RMSEA (90% CI)
Gender invariance											
(9) Configural invariance	–	5,502.98	–	1,294	–	–	0.912	–	0.904	–	0.054 (0.052–0.055)
(10) Metric invariance	–	5,542.01	–	1,326	–	–	0.912	–	0.906	–	0.053 (0.052–0.055)
	10 versus 9	–	39.03	–	32	N.S.	–	<0.001	–	Yes	–
(11) Scalar invariance	–	5,729.31	–	1,364	–	–	0.908	–	0.906	–	0.053 (0.052–0.055)
	11 versus 10	–	187.30	–	38	$p < 0.001$	–	0.004	–	Yes	–
(12) Error variance invariance	–	6,229.12	–	1,405	–	–	0.899	–	0.899	–	0.055 (0.054–0.057)
	12 versus 11	–	499.81	–	41	$p < 0.001$	–	0.009	–	Yes	–
Subgroup invariance											
(13) Configural invariance	–	5,520.98	–	1,294	–	–	0.912	–	0.904	–	0.054 (0.053–0.055)
(14) Metric invariance	–	5,567.64	–	1,326	–	–	0.912	–	0.906	–	0.053 (0.052–0.055)
	14 versus 13	–	46.66	–	32	$p < 0.05$	–	<0.001	–	Yes	–
(15) Scalar invariance	–	5,608.20	–	1,364	–	–	0.912	–	0.909	–	0.053 (0.051–0.054)
	15 versus 14	–	40.56	–	38	N.S.	–	<0.001	–	Yes	–
(16) Error variance invariance	–	5,710.42	–	1,405	–	–	0.910	–	0.910	–	0.052 (0.051–0.054)
	16 versus 15	–	102.22	–	41	$p < 0.001$	–	0.002	–	Yes	–

N_{whole} = 2,240; N_{males} = 742; $N_{females}$ = 1,498; N_{odd} = 1,120; N_{even} = 1,120; CFI, Comparative Fit Index; RMSEA, Root Mean Square Error of Approximation; CI, confidence interval; $\Delta\chi^2$, change in χ^2 compared to the previous model; Δdf, change in degrees of freedom compared to the previous model; N.S., $\Delta\chi^2$ not significant at $p < 0.05$; ΔCFI, change in CFI compared to the previous model; ΔCFI ≤ |0.01|?, Was the change in CFI not larger than the |0.01|-cutoff?; Model 9 and Model 13, no equality constraints were imposed; Model 10 and Model 14, equality constraints were imposed on all factor loadings; Model 11 and Model 15, equality constraints were imposed on all factor loadings and intercepts of the measured variables; Model 12 and Model 16, equality constraints were imposed on all factor loadings, intercepts, and residual variances.

employers looking for candidates possessing key behavioral attributes of an effective service leader. Finally, the developed tool can help researchers to conduct studies on service leadership in the changing service economy in the global context.

While the present study is pioneer in the area of service leadership, there are several limitations of the study. First, only undergraduate students in Hong Kong were recruited in the present study. Hence, it would be helpful to understand the psychometric properties of the measure in other student populations. Besides, to further endorse the factorial validity of the SLB-SF-38, follow-up validation studies using a sample of executives (e.g., Acar and Zehir, 2009) or managers (e.g., Yukl et al., 2008) are suggested.

Second, given that the present survey comprised over 250 items, response burden may influence the response quality (Lavrakas, 2008). Besides, content overlap could also be a "turn-off" for the respondents (Rolstad et al., 2011). In addition,

TABLE 10 | Correlation coefficients, mean inter-item correlations and Cronbach's alpha amongst the six subscales and the whole scale.

		Subscales					
Subscales	**Cronbach's alpha (Mean inter-item correlations)**	**1**	**2**	**3**	**4**	**5**	**6**
(1) Self-improvement and Self-reflection	0.909 (0.530)	–	–	–	–	–	–
(2) People and Principles Orientation	0.868 (0.426)	0.70	–	–	–	–	–
(3) Resilience	0.880 (0.515)	0.61	0.57	–	–	–	–
(4) Social Competence	0.868 (0.570)	0.61	0.69	0.64	–	–	–
(5) Problem-Solving	0.853 (0.538)	0.58	0.50	0.61	0.52	–	–
(6) Mentorship	0.847 (0.647)	0.58	0.58	0.50	0.49	0.42	–
38-item Service Leadership Behavior Scale	0.958 (0.377)	0.87	0.85	0.82	0.81	0.74	0.70

N = 2,240. All correlation coefficients are statistically significant at $p < 0.001$ (two-tailed).

TABLE 11 | Correlations with external criterion scales (and subscales).

	External criterion scales (and subscales)					
	RSLP	**MSC**	**LEF**	**IRI**	**IRI-EC**	**IRI-PT**
38-item Service Leadership Behavior Scale	0.79	0.66	0.52	0.44	0.31	0.47
Subscale 1: Self-improvement and Self-reflection	0.69	0.58	0.45	0.38	0.28	0.40
Subscale 2: People and Principles Orientation	0.78	0.70	0.37	0.55	0.44	0.53
Subscale 3: Resilience	0.57	0.44	0.42	0.23	0.12	0.30
Subscale 4: Social Competence	0.62	0.58	0.46	0.42	0.32	0.43
Subscale 5: Problem-Solving	0.49	0.37	0.45	0.20	0.10	0.26
Subscale 6: Mentorship	0.64	0.45	0.40	0.25	0.16	0.29

N = 2,240. All correlation coefficients are statistically significant at $p < 0.001$ (two-tailed). RSLP, Revised Servant Leadership Profile; MSC, Moral Self-Concept; LEF, Leadership Efficacy; IRI, Interpersonal Reactivity Index; IRI-EC, Subscale "Empathic Concern"; IRI: PT, Subscale "Perspective Taking."

TABLE 12 | Correlations with other Service Leadership scales (and subscales) under validation.

	SLK-SF-40	SLA-F1	SLA-F2	SLA-F3	SLA-F4	SLA-F5	SLA-F6	SLA-F7	SLA-F8	SLA-SF-46
38-item Service Leadership Behavior Scale	0.19	0.51	0.49	0.51	0.40	0.50	0.47	0.28	−0.05	0.58
Subscale 1: Self-improvement and Self-reflection	0.20	0.46	0.42	0.42	0.33	0.44	0.43	0.20	$-0.03^{n.s.}$	0.50
Subscale 2: People and Principles Orientation	0.28	0.52	0.55	0.52	0.41	0.56	0.47	0.28	$0.02^{n.s.}$	0.62
Subscale 3: Resilience	0.06	0.35	0.33	0.37	0.29	0.32	0.33	0.23	−0.09	0.39
Subscale 4: Social Competence	0.20	0.42	0.43	0.41	0.31	0.44	0.38	0.23	$0.02^{n.s.}$	0.49
Subscale 5: Problem-Solving	0.12	0.38	0.34	0.34	0.25	0.33	0.36	0.13	−0.05	0.39
Subscale 6: Mentorship	−0.08	0.26	0.26	0.40	0.35	0.22	0.22	0.27	−0.18	0.32

N = 2,240. Unless otherwise specified by superscript "n.s." which denotes statistical non-significance, all correlation coefficients are significant at $p < 0.05$ (two-tailed). SLK-SF-40, Scale score of the one-factor, 40-item Service Leadership Knowledge Scale; SLA-SF-46, Scale score of the eight-factor, 46-item Service Leadership Attitude Scale; SLA-F1, Factor "Vision and competence"; SLA-F2, Factor "People orientation"; SLA-F3, Factor "Caring disposition"; SLA-F4, Factor "Ethical role model"; SLA-F5, Factor "Social competence"; SLA-F6, Factor "Self-understanding and reflection"; SLA-F7, Factor "Positive view about human beings"; SLA-F8, Factor 8 "Unchangeable and dark human nature."

although findings provide strong support for the internal consistency of the SLB, the test-retest reliability analyses can be conducted to examine the temporal stability of the measure in future. Nevertheless, our results showed good internal consistency of both the scale and the subscales (see **Table 4**), implying the quality responses from the participants (Oltedal et al., 2007).

Third, the SLB-SF-38 relies on participants' self-rated leadership behavior, which may cause social desirability bias in responses. Participants may tend to provide favorable instead of truthful responses. Although we assured the participants that the responses would be kept confidential and anonymous, this limitation should be taken into account. In future, additional information collected from other informants (e.g., followers) would give a more comprehensive picture about service leadership behavior seen from different perspectives.

Finally, one can criticize that because the data are ordinal data, it is not appropriate to use parametric factor analysis. While we acknowledge this weakness of the present paper, we would like to make several arguments supporting the approach adopted in this study. Primarily, although there are contrary views, it is a common practice to treat ordinal data with several response categories as continuous data (Muthén and Kaplan, 1985). Second, it is also a common practice to apply CFA with ML estimation to test the model of Likert scale measurement (Byrne, 2010). For example, similar papers using CFA to analyze Likert scale data have been reported in some prestigious journals, including *Frontiers in Psychology* and *Psychological Assessment* (Young and Beaujean, 2011; Coates et al., 2016; Jorge-Monteiro and Ornelas, 2016; Ghislieri et al., 2017).

Third, Carifio and Perla discussed some common misunderstandings about Likert scales and regarded the claim that "because Likert scales are ordinal-level scales, only non-parametric statistical tests should be used with them" (Carifio and Perla, 2007, p. 114) as a common myth. They further pointed out that "if one is using a 5–7 point Likert response format, and particularly so for items that resemble a Likert-like scale and factorially hold together as a scale or subscale reasonably well, then it is perfectly acceptable and correct to analyze the results at the (measurement) scale level using parametric analyses techniques such as the F-Ratio or the Pearson correlation coefficients or its extensions (i.e., multiple regression and so on), and the results of these analyses should and will be interpretable as well" (Carifio and Perla, 2007, p. 115).

Fourth, we understand that other estimators (e.g., WLSMV) can be superior to ML when there are few ordinal categories. However, there are views supporting the application of ML for categorical data under specific conditions (Byrne, 2010). Some researchers have compared ML and other estimators applied for CFA analysis with ordered categorical data, such as WLSMV (Beauducel and Herzberg, 2006), WLS (Lei, 2009), GLS (Muthén and Kaplan, 1985; Hu and Bentler, 1998), and cat-LS (Rhemtulla et al., 2012). Most of these comparisons concluded that ML performed as good as or even better than other methods when (a) the data approximated a normal distribution (have mildly to moderately skewed/kurtosis variables), (b) there were more than five response categories, and (c) the sample size was not small. In this study, these three conditions were fully met. On the other hand, some researchers have highlighted the disadvantages of WLSMV. For example, Li pointed out the weaknesses of inter factor correlations and standard errors in WLSMV estimation "when the sample size is small, and/or when a latent distribution is moderately non-normal" (Li, 2016, p. 948). In addition, DiStefano and Morgan (2014) also noticed that WLSMV may produce factor correlation estimates with overestimation when dealing with five or more ordered categories.

Finally, as suggested by Rhemtulla et al. (2012), the choice of available methods should rely on data characters (e.g., sample size, model size, the normality of distribution), the characters of constructs underlying (e.g., the distribution of the constructs), and researchers' own interests. In the present study, the data in general showed a normal distribution, the sample size was relatively large, and six response categories were used. In this regard, ML seems appropriate. As suggested by Allison et al. (1993, p. 92) recommended researchers "should consider staying with traditional parametric tests" when the above conditions are met. Obviously, ML provides better robust standard errors for factor correlations and the desirable asymptotic properties such as asymptotically efficiency (Lei, 2009; Rhemtulla et al., 2012).

In short, we understand the reviewer's concern. We acknowledge the related limitations of the study and we suggest a future study to be conducted to provide an additional picture. Despite this limitation, the present study provides pioneer and exciting support for a pioneer scale on service leadership behavior in a Chinese context.

CONCLUSION

Despite the above limitations, the present study provides evidence for a reliable and valid assessment tool of service leadership behavior. The present analyses provide a strong evidence base for the psychometric properties of the SLB-SF-38 by using a large sample of Chinese undergraduates. The current study fills the gap in the scientific literature on leadership assessment of leadership training amongst Chinese college students, and also provides practical implications for future service leadership education and research.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was approved by the Human Subjects Ethics Sub-committee (HSESC) (or its Delegate) of The Hong Kong Polytechnic University. All subjects have given written informed consent before start of the study.

AUTHOR CONTRIBUTIONS

DS designed the research project and contributed to all the steps of the work. DD contributed to the development of the article and revised the manuscript based on the critical comments and editing provided by DS. LM contributed to the initial data analyses and development of a rough draft of the manuscript.

FUNDING

The Fung Service Leadership Education Initiative (FSLEI) and this work were financially supported by the Victor and William Fung Foundation and the Endowed Professorship in Service Leadership Education at The Hong Kong Polytechnic University.

REFERENCES

Acar, A. Z., and Zehir, C. (2009). Development and validation of a multidimensional business capabilities measurement instrument. *J. Transnatl. Manag.* 14, 215–240. doi: 10.1080/15475770903127050

Allison, D. B., Gorman, B. S., and Primavera, L. H. (1993). Some of the most common questions asked of statistical consultants: our favorite responses and recommended readings. *J. Group Psychother. Psychodrama Sociom.* 46, 83–109.

Altun, I. (2003). The perceived problem-solving ability and values of student nurses and midwives. *Nurse Educ. Today* 23, 575–584. doi: 10.1016/s0260-6917(03)00096-0

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychol. Bull.* 103, 411–423.

Aquino, K., and Reed, A. (2002). The self-importance of moral identity. *J. Personal. Soc. Psychol.* 83, 1423–1440.

Avolio, B. J., Bass, B. M., and Jung, D. I. (1999). Re-examining the components of transformational and transactional leadership using the multifactor Leadership. *J. Occup. Organ. Psychol.* 72, 441–462. doi: 10.1348/096317999166789

Awang, Z. (2012). *A Handbook on SEM: Structural Equation Modelling Using AMOS Graphics*, 2nd Edn. Malaysia: Universiti Sultan Zainal Abidin press.

Bacon, C., Benton, D., and Gruneberg, M. M. (1979). Employers' opinions of university and polytechnic graduates. *Vocat. Aspect Educ.* 31, 95–102. doi: 10.1080/10408347308001251

Bass, B. M. (1990). From transactional to transformational leadership: learning to share the vision. *Organ. Dyn.* 18, 19–31. doi: 10.1016/0090-2616(90)90061-s

Beauducel, A., and Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Struct. Equ. Mod.* 13, 186–203. doi: 10.1207/s15328007sem1302_2

Besnoy, K. D., Dantzler, J., Besnoy, L. R., and Byrne, C. (2016). Using exploratory and confirmatory factor analysis to measure construct validity of the traits, aptitudes, and Behaviors Scale (TABS). *J. Educ. Gift.* 39, 3–22. doi: 10.1177/0162353215624160

Biau, D. J., Kerneis, S., and Porcher, R. (2008). Statistics in brief: the importance of sample size in the planning and interpretation of medical research. *Clin. Orthop. and Relat. Res.* 466, 2282–2288. doi: 10.1007/s11999-008-0346-9

Brown, M. E., and Treviño, L. K. (2006). Ethical leadership: a review and future directions. *Leadersh. Q.* 17, 595–616. doi: 10.1016/j.leaqua.2006.10.004

Bryson, J. R., and Daniels, P. W. (2015). *Handbook of Service Business Management, Marketing, Innovation and Internationalisation*, Eds. Edn. Cheltenham: Edward Elgar Publishing.

Byrne, B. M. (1998). *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming*, 2nd Edn. New York, NY: Routledge.

Carifio, J., and Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *J. Soc. Sci.* 3, 106–116. doi: 10.1016/0006-8993(93)90283-S

Cheng, C. H. K. (2005). *The Chinese Adolescent Self-Esteem Scales (CASES): A User Manual*. Hong Kong: City University of Hong Kong Press.

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Mod.* 9, 233–255. doi: 10.1097/NNR.0b013e3182544750

Chou, C.-P., and Bentler, P. M. (1995). "Estimates and tests in structural equation modeling," in *Structural equation modeling: Concepts, Issues and Applications*, ed. R. H. Hoyle (Thousand Oaks, CA: Sage), 37–55.

Chung, P. P. Y. (2015). "Where there is no vision, the people will perish," in *Promoting Service Leadership Qualities In university Students: The case of Hong Kong* (pp. xv-xviii), eds D. T. L. Shek and P. P. Y. Chung (Singapore: Springer).

Chung, P. P. Y., and Bell, A. H. (2015). *25 Principles of Service Leadership*, 1st Edn. New York, NY: Lexingford Publishing.

Chung, P. P. Y., and Elfassy, R. (2016). *The 12 Dimensions of a Service Leader*, 1st Edn. New York, NY: Lexingford Publishing.

Coates, R., Ayers, S., and de Visser, R. (2016). Factor structure of the edinburgh postnatal depression scale in a population-based sample. *Psychol. Assess.* 29, 1016–1027. doi: 10.1037/pas0000397

Davis, M. H. (1983). Measuring individual-differences in empathy: evidence for a multidimensional approach. *J. Personal. Soc. Psychol.* 44, 113–126. doi: 10.1037//0022-3514.44.1.113

DiStefano, C., and Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust rstimation techniques for ordinal data. *Struct. Equ. Mod.* 21, 425–438. doi: 10.1080/10705511.2014.915373

Ghislieri, C., Emanuel, F., Molino, M., Cortese, C. G., and Colombo, L. (2017). New technologies smart, or harm work-family boundaries management? Gender differences in conflict and enrichment using the JD-R theory. *Front. Psychol.* 8:1070. doi: 10.3389/fpsyg.2017.01070

Greenleaf, R. K. (1970). *The Servant as a Leader*. Indianapolis, IN: Greenleaf Center.

Greenleaf, R. K. (1977). *Servant Leadership: A Journey into the Nature of Legitimate Power and Greatness*. Mahwah, NJ: Paulist Press.

Hatler, C., and Sturgeon, P. (2013). Resilience building: a necessary leadership competence. *Nurse Lead.* 11, 32–39. doi: 10.1016/j.mnl.2013.05.007

Hirsh, J. B. (2010). Personality and environmental concern. *J. Environ. Psychol.* 30, 245–248. doi: 10.1016/j.jenvp.2010.01.004

Ho, J., and Nesbit, P. L. (2009). A refinement and extension of the self-leadership scale for the Chinese context. *J. Manag. Psychol.* 24, 450–476. doi: 10.1108/02683940910959771

Hong, Y., Liao, H., Hu, J., and Jiang, K. (2013). Missing link in the service profit chain: a meta-analytic review of the antecedents, consequences, and moderators of service climate. *J. Appl. Psychol.* 98, 237–267. doi: 10.1037/a0031666

Hong Kong Institute of Service Leadership and Management Limited [HKI-SLAM] (2013). *An Overview of HKI-SLAM's Curriculum Framework Prepared for Li & Fung's Service Leadership Initiative*. Available at: http://hki-slam.org/files/press/An%20Overview%20of%20SLAM%20Curriculum%20Framework%20130531.pdf (accessed August 20, 2017).

Hu, L., and Bentler, P. M. (1998). Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychol. Methods* 3, 424–453. doi: 10.1037/1082-989X.3.4.424

Jasovsky, D. A., and Kamienski, M. (2007). "Enhancing your critical thinking, decision making, and problem solving," in *Nursing Leadership and Management: Theories, processes and practice* (1st Edn, ed. R. P. Jones (Philadelphia, PA: FA Davis), 151–165.

Jiang, K., Chuang, C.-H., and Chiao, Y.-C. (2015). Developing collective customer knowledge and service climate: the interaction between service-oriented high-performance work systems and service leadership. *J. Appl. Psychol.* 100, 1089–1106. doi: 10.1037/apl0000005

Jiang, K., Hu, J., Hong, Y., Liao, H., and Liu, S. (2016). Do it well and do it right: the impact of service climate and ethical climate on business performance and the boundary conditions. *J. Appl. Psychol.* 101, 1553–1568. doi: 10.1037/apl0000138

Jorge-Monteiro, M. F., and Ornelas, J. H. (2016). Recovery assessment scale: testing validity with portuguese community-based mental health organization users. *Psychol. Assess.* 28, 1–11. doi: 10.1037/pas0000176

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*, 2nd Edn. New York: The Guilford Press.

Kopelman, R. E., Prottas, D. J., and Davis, A. L. (2008). Douglas McGregor's theory X and Y: toward a construct-valid measure. *J. Manag. Issues* 20, 255–271.

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: SAGE.

Lei, P. W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Q. Q.* 43, 495–507. doi: 10.1007/s11135-007-9133-z

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behav. Res. Methods* 48, 936–949. doi: 10.3758/s13428-015-0619-7

Lytle, R. S., Hom, P. W., and Mokwa, M. P. (1998). SERV*OR: a managerial measure of organizational service-orientation. *J. Retailing* 74, 455–489. doi: 10.1016/s0022-4359(99)80104-3

Ma, C. M., Shek, D. T., and Chandra, Y. (2018). Development of the attitude to service leadership scale in hong kong [Special issue]. *Int. J. Child Adolesc. Health* 11, 405–414.

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychol. Methods* 4, 84–99.

Mehrdad, A., Farhad, S., and Maryam, B. (2011). The effects of problem solving skills training on test anxiety among college students. *Dev. Psychol. J. Iran. Psychol.* 8, 67–74.

Mendonca, M. (2001). Preparing for ethical leadership in organizations. *Can. J. Administr. Sci.* 18, 266–276. doi: 10.1111/j.1936-4490.2001.tb00262.x

Milfont, T. L., and Fischer, R. (2010). Testing measurement invariance across groups: applications in crosscultural research. *Int. J. Psychol. Res.* 3, 111–121.

Mumford, M. D., Zaccaro, S. J., Harding, F. D., Jacobs, T. O., and Fleishman, E. A. (2000). Leadership skills for a changing world: solving complex social problems. *Leadersh. Q.* 11, 11–35.

Murphy, S. E. (1992). *The Contribution of Leadership Experience and Self-Efficacy to Group Performance Under Evaluation Apprehension*. Washington: University of Washington.

Muthén, B., and Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *Br. J. Math. Statist. Psychol.* 38, 171–189. doi: 10.1111/j.2044-8317.1985.tb00832.x

Muthén, L. K., and Muthén, B. O. (1998–2010). *Mplus Use"s Guide*, 6th Edn. Los Angeles, CA: Muthén & Muthén. doi: 10.1111/j.2044-8317.1985.tb00832.x

Oltedal, S., Garratt, A., Bjertns, O., Bjornsdottir, M., Freil, M., and Sachs, M. (2007). The NORPEQ patient experiences questionnaire: data quality, internal consistency and validity following a Norwegian inpatient survey. *Scand. J. Public Health* 35, 540–547. doi: 10.1080/14034940701291724

Orpinas, P. (2010). "Social competence," in *The Corsini Encyclopedia of Psychology*, 4th Edn, Vol. 4, eds I. B. Weiner and W. E. Craighead (Hoboken, NJ: Wiley), 1–2.

Page, D., and Wong, P. T. P. (2000). "A conceptual framework for measuring Servant Leadership," in *The Human Factor in Shaping the Course of History and Development*, ed. S. B. S. K. Adjibolooso (Lanham, MD: University Press of America), 69–109.

Park, G.-P. (2014). Factor analysis of the foreign language classroom anxiety scale in korean learners of english as a foreign language. *Psychol. Rep.* 115, 261–275. doi: 10.2466/28.11.PR0.115c10z2

Patel, B. (2012). *The Importance of Resilience in Leadership*. London: Clore Social Leadership.

Rhemtulla, M., Brosseau-Liard, P. É, and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychol. Methods* 17, 354–373. doi: 10.1037/a0029315

Rolstad, S., Adler, J., and Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value Health* 14, 1101–1108. doi: 10.1016/j.jval.2011.06.003

Russell, R. F., and Stone, A. G. (2002). A review of servant leadership attributes: Developing a practical model. *Leadersh. Organ. Dev. J.* 23, 145–157. doi: 10.1108/01437730210424

Schmitt, N., and Kuljanin, G. (2008). Measurement invariance: review of practice and implications. *Hum. Res. Manag. Rev.* 18, 210–222. doi: 10.1016/j.hrmr.2008.03.003

Schneider, B., Ehrhart, M. G., Mayer, D. M., Saltz, J. L., and Niles-Jolly, K. (2005). Understanding organization-customer links in service settings. *Acad. Manag. J.* 48, 1017–1032. doi: 10.5465/amj.2005.19573107

Schneider, B., White, S. S., and Paul, M. C. (1998). Linking service climate and customer perceptions of service quality: test of a causal model. *J. Appl. Psychol.* 83, 150–163. doi: 10.1037/0021-9010.83.2.150

Sendjaya, S., and Sarros, J. C. (2002). Servant leadership: Its origin, development, and application in organizations. *J. Leadersh. Organ. Stud.* 9, 57–64. doi: 10.1177/107179190200900205

Shek, D. T. L., and Chai, W. Y. (2019). Psychometric properties of the service leadership attitude scale in hong kong. *Front. Psychol.* 10:1070. doi: 10.3389/fpsyg.2019.01070

Shek, D. T. L., and Chung, P. P. Y. (eds) (2015). *Promoting Service Leadership Qualities in University Students* (1st Edn.). Singapore: Springer.

Shek, D. T. L., Chung, P. P. Y., and Leung, H. (2015a). How unique is the service leadership model? A comparison with contemporary leadership approaches. *Int. Disabil. Hum. Dev.* 14, 217–231.

Shek, D. T. L., Chung, P. P. Y., and Leung, H. (2015b). Manufacturing economy vs. service economy: implications for service leadership. *Int. J. Disabil. Hum. Dev.* 14, 205–215.

Shek, D. T. L., Chung, P. P. Y., Lin, L., Leung, H., and Ng, E. C. W. (2018a). "Service Leadership under the Service Economy," in *Global and Culturally Diverse Leaders and Leadership*, 1st Edn, eds J. L. Chin, J. E. Trimble, and J. E. Garcia (Bingley: Emerald Publishing), 143–161.

Shek, D. T. L., Ma, L. K., Lin, L., and Leung, H. (2018b). Psychometric properties of the service leadership behavior scale: preliminary findings. *Int. J. Child Adolesc. Health* 11, 427–443.

Shek, D. T. L., Ma, L. K., Ma, M. S. C., and Hoshmand, R. A. (2018c). Convergent and factorial validation of the service leadership behavior scale [Special issue]. *Int. J. Child Adolesc. Health* 11, 479–492.

Shek, D. T. L., Ma, L. K., Yu, L., and Leung, L. M. (2018d). Validation of the service leadership knowledge scale: factorial and convergent validity [Special issue]. *Int. J. Child Adolesc. Health* 11, 455–466.

Shek, D. T. L., Zhu, X., and Chan, K.-M. (2018e). Development of service leadership behavior scale: background and conceptual model [Special issue]. *Int. J. Child Adolesc. Health* 11, 415–424.

Shek, D. T. L., Zhu, Y. F. A., Ma, K. L., and Lin, L. (2018f). Validation of the service leadership attitude scale in hong kong [Special issue]. *Int. J. Child Adolesc. Health* 11, 467–477. doi: 10.3389/fpsyg.2019.01070

Shek, D. T. L., Chung, P. P. Y., Lin, L., and Merrick, J. (2017). *Service Leadership Education for University Students*, Eds Edn. New York, NY: Nova Science.

Shek, D. T. L., Chung, P. P. Y., Yu, L., and Merrick, J. (2015c). Service leadership curriculum and higher education reform in hong kong [Special issue]. *Int. J. Disabil. Hum. Dev.* 14, 297–306.

Shek, D. T. L., and Leung, H. (2015). "Service Leadership qualities in university students through the lens of student well-being," in *Promoting Service Leadership Qualities in University Students* (1st Edn, eds D. T. L. Shek and P. P. Y. Chung (Singapore: Springer), 1–16. doi: 10.1007/978-981-287-515-0_1

Shek, D. T. L., and Li, X. (2015). The role of a caring disposition in Service Leadership. *Int. J. Disabil. Hum. Dev.* 14, 319–332.

Shek, D. T. L., and Lin, L. (2015a). Core beliefs in the service leadership model proposed by the hong kong institute of service leadership and management. *Int. J. Disabil. Hum. Dev.* 14, 233–242.

Shek, D. T. L., and Lin, L. (2015b). "Evaluating Service Leadership programs with multiple strategies," in *Promoting Service Leadership Qualities in University Students* (1st Edn, eds D. T. L. Shek and P. P. Y. Chung (Singapore: Springer), 197–211. doi: 10.1007/978-981-287-515-0_13

Shek, D. T. L., and Lin, L. (2015c). Intrapersonal competencies and service leadership. *Int. J. Disabil. Hum. Dev.* 14, 255–263.

Shek, D. T. L., and Lin, L. (2015d). Leadership and mentorship: service leaders as mentors of the followers. *Int. J. Disabil. Hum. Dev.* 14, 351–359.

Shek, D. T. L., and Lin, L. (2017). "Validation of the Service Leadership Knowledge Scale: Criterion-related validity," in *Service Leadership Education for University Students*, Eds. Edn, eds D. T. L. Shek, P. P. Y. Chung, L. Lin, and J. Merrick (New York NY: Nova Science), 189–204.

Shek, D. T. L., and Ma, C. M. S. (2010). Dimensionality of the chinese positive youth development scale: confirmatory factor analyses. *Soc. Indic. Res.* 98, 41–59. doi: 10.1007/s11205-009-9515-9

Shek, D. T. L., and Ma, C. M. S. (2014). Validation of a subjective outcome evaluation tool for participants in a positive youth development program in hong kong. *J. Pediatr. Adolesc. Gynecol.* 27(Suppl.), S43–S49.

Shek, D. T. L., and Yu, L. (2014). Factorial validity of a subjective outcome evaluation tool for implementers of a positive youth development program. *J. Pediatr. Adolesc. Gynecol.* 27, S32–S42. doi: 10.1016/j.jpag.2014.02.010

Snell, R. S., Chan, M. Y. L., and Zou, T. X. P. (2017). "Key practices of leadership for service in Hong Kong," in *Service Leadership Education for University Students*, eds D. T. L. Shek, P. P. Y. Chung, L. Lin, and J. Merrick (New York, NY: Nova Science), 127–138.

Swami, V., Barron, D., Weis, L., Voracek, M., Stieger, S., and Furnham, A. (2017). An examination of the factorial and convergent validity of four measures of conspiracist ideation, with recommendations for researchers. *PLoS One* 12:e0172617. doi: 10.1371/journal.pone.0172617

Towers Watson (2012). *The Next High-Stakes Quest: Balancing Employer and Employee Priorities— 2012-2013 Global Talent Management and Rewards Study*. Available at: https://www.towerswatson.com/en-HK/Insights/IC-Types/Survey-Research/Results/2012/09/2012-Global-Talent-Management-and-Rewards-Study (accessed April 21, 2018).

University Grants Committee [UGC]. (2017). *Student Enrolment of UGC-funded Programmes by University, Level of Study, Mode of Study and Sex, 2010/11 to 2016/17*. Hong Kong: UGC.

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Wielkiewicz, R. M. (2000). The leadership attitudes and beliefs scale: an instrument for evaluating college students' thinking about leadership and organizations. *J. Coll. Stud. Dev.* 41, 335–347.

Wong, A., Liu, Y., and Tjosvold, D. (2015). Service leadership for adaptive selling and effective customer service teams. *Ind. Mark. Manag.* 46, 122–131. doi: 10.1016/j.indmarman.2015.01.012

Wong, P. T. P., and Page, D. (2003). *Servant Leadership: An Opponent-Process Model and the Revised Servant Leadership Profile*. Available at: https://www.regent.edu/acad/global/publications/sl_proceedings/2003/wong_servant_leadership.pdf (accessed July 31, 2018).

Wu, H.-C., and Mohi, Z. (2015). Assessment of service quality in the fast-food restaurant. *J. Foodserv. Bus. Res.* 18, 358–388. doi: 10.1080/15378020.2015.1068673

Young, J. K., and Beaujean, A. A. (2011). Measuring personality in wave I of the national longitudinal study of adolescent health. *Front. Psychol.* 2:158. doi: 10.3389/fpsyg.2011.00158

Yukl, G., Seifert, C. F., and Chavez, C. (2008). Validation of the extended influence behavior questionnaire. *Leadersh. Q.* 19, 609–621. doi: 10.1016/j.leaqua.2008.07.006

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 07 August 2019
doi: 10.3389/fpsyg.2019.01841

Assessing Callous-Unemotional Traits in Chinese Detained Boys: Factor Structure and Construct Validity of the Inventory of Callous-Unemotional Traits

Xintong Zhang[1,2], Yiyun Shou[3], Meng-Cheng Wang[1,2,4], Chuxian Zhong[1,2], Jie Luo[5], Yu Gao[6] and Wendeng Yang[1,4]*

[1] Department of Psychology, Guangzhou University, Guangzhou, China, [2] The Center for Psychometrics and Latent Variable Modeling, Guangzhou University, Guangzhou, China, [3] Research School of Psychology, The Australian National University, Canberra, ACT, Australia, [4] The Key Laboratory for Juveniles Mental Health and Educational Neuroscience in Guangdong Province, Guangzhou University, Guangzhou, China, [5] School of Psychology, Guizhou Normal University, Guiyang, China, [6] Brooklyn College, The City University of New York, New York, NY, United States

OPEN ACCESS

Edited by:
Elisa Pedroli,
Italian Institute for Auxology (IRCCS),
Italy

Reviewed by:
Geert Jan Stams,
University of Amsterdam, Netherlands
Matt DeLisi,
Iowa State University, United States

****Correspondence:***
Meng-Cheng Wang
wmcheng2006@126.com

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *12 April 2019*
Accepted: *25 July 2019*
Published: *07 August 2019*

Citation:
Zhang X, Shou Y, Wang M-C, Zhong C, Luo J, Gao Y and Yang W (2019) Assessing Callous-Unemotional Traits in Chinese Detained Boys: Factor Structure and Construct Validity of the Inventory of Callous-Unemotional Traits. Front. Psychol. 10:1841. doi: 10.3389/fpsyg.2019.01841

The Inventory of Callous-Unemotional Traits (ICU) was designed to evaluate multiple facets of Callous-Unemotional (CU) traits in youths. However, no study has examined the factor structure and psychometrical properties of the ICU in Chinese detained juveniles. The current study assesses the factor structure, internal consistency and convergent validity of the ICU in 613 Chinese detained boys. Confirmatory factor analysis results indicated that the original three-factor model with 24 items showed an unacceptable fit to the data, however, the 11-item shortened version of the ICU (ICU-11) with callousness and uncaring dimensions showed the best fit. Moreover, the ICU-11 total score and factor scores had good and acceptable internal consistencies. The convergent and criterion validity of the ICU-11 was demonstrated by comparable and significant associations in the expected direction with relevant external criteria (e.g., psychopathy, aggression, and empathy). In conclusion, present findings indicated that the ICU-11 is a reliable and efficient instrument to replace the original ICU when assessing CU traits in the Chinese male detained juvenile sample.

Keywords: callous-unemotional traits, psychopathy, detained juvenile, factor structure, confirmatory factor analysis, validation

INTRODUCTION

The Callous-Unemotional (CU) traits in children and adolescents are a specifier of the criteria for conduct disorder (CD) in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5, American Psychiatric Association, 2013), and are considered as an affective characteristic of psychopathic personality disorder (Frick and Moffitt, 2010). And the CU traits have been proven to be the most crucial predictors of criminal activities (Asscher et al., 2011). Features of a high level of the CU traits include a lack of concern about performance, shallow emotions, a lack of empathy and guilt, and having low sensitivity to others' feelings (Frick, 2009). As such, the

CU traits may be used to define a subgroup of youths with severe and persistent conduct problems, delinquency, or aggression particularly referring to a more proactive type of aggression (Kahn et al., 2012; Byrd et al., 2013). Different from other antisocial juveniles, those with CU traits tend to have difficulty in dealing with negative emotional stimuli (Kimonis et al., 2008), a lack of fearful inhibitions and anxiety (Frick et al., 1999) and a lack of sensitivity to punishment cues (Fisher and Blair, 1998). Remarkably, psychopathy is one of the most important predictors of criminality (DeLisi and Vaughn, 2015; DeLisi, 2016; DeLisi et al., 2018). Substantial evidence has demonstrated that the juvenile with higher psychopathy especially those have affective deficits and less self-control, had increased likelihood of engaging in violent forms of antisocial behaviors (DeLisi et al., 2010, 2018), in criminal careers that continue into the adulthood (Vaughn and DeLisi, 2008).

Understanding CU traits in delinquent and antisocial adolescents requires efficient, reliable and valid measurement tools. The Inventory of Callous-Unemotional Traits (ICU) was developed as a stand-alone and comprehensive self-report instrument (Frick, 2004). The ICU contains 24 items that are expanded from the CU factor (four items) of the Antisocial Process Screening Device (APSD; Frick and Hare, 2001). Since its introduction, various informant versions of the ICU have been increasingly endorsed in research, and have demonstrated reliable associations with external criteria variables in both incarcerated and community youth (Roose et al., 2010; Pihet et al., 2015; Pechorro et al., 2016b, 2017). However, a recent meta-analysis by Deng et al. (2019) has noted that there remains a lack of evidence of the applicability of the ICU among non-European-American samples. Although there has been an attempt of validating the ICU among Chinese community samples (Wang et al., 2017b, 2019), little is known of the utility of the ICU in clinical settings in non-English-speaking delinquent populations.

Furthermore, although the ICU was originally developed as a unidimensional measure of CU traits (an overarching CU factor containing three subfactors: unemotional, callousness and uncaring), this early proposed three-factor, as well as a three-factor bifactor model (Essau et al., 2006), received limited support in either community (Ciucci et al., 2014; Wang et al., 2017b, 2019) or delinquent samples (e.g., Kimonis et al., 2008) due to the poor overall fit of these models. Notably, the unemotional factor has been shown to have relatively poor psychometric properties, showing low reliability, poor factor loadings and inadequate correlations with external criteria (e.g., Essau et al., 2006; Kimonis et al., 2008; Byrd et al., 2013). Many recent studies have excluded some or all of the unemotional factor items, and have focused on developing a range of short versions of the ICU.

For example, Hawes et al. (2014) developed a 12-item shortened form of the ICU (ICU-12) using item response theory. The ICU-12 has two correlated factors: callousness (seven items) and uncaring (five items), and its validity and reliability were supported in a number of subsequent studies that used detained samples (e.g., Colins et al., 2016; Paiva-Salisbury et al., 2017). Two recent studies found that an 11-item model (ICU-11) which excluded the item, “I do not show my emotions to others” – the only item retained from the unemotional factor – achieved a better fit than the ICU-12 among Chinese-speaking samples using university students (Wang et al., 2017b) and community children (Wang et al., 2019). This is possibly due to the fact that expressing emotion is generally not encouraged in Chinese culture, thus resulting in the low discriminability of the item among Chinese populations. Nevertheless, the ICU-11 displayed measurement invariance across informants and occasions and had strong evidence for its criteria validity (Wang et al., 2019). The results of Wang et al. (2017b) also showed strong associations with other measures of psychopathic traits, and both of the two factors (callousness and uncaring) correlated significantly with the total scores on the ASPD and proactive aggression.

Psychopathy has been integrated into mainstream criminological theories (DeLisi and Vaughn, 2015), and at least in part, explains the causal mechanisms underlying chronic, serious, and violent delinquent trajectories, so that psychopathy can be used as a risk for the development and maintenance of delinquent behaviors (Asscher et al., 2011; Corrado et al., 2015). Moreover, regardless the intensity of the violence, the CU traits were found significantly correlated with violent offending (Sherretts et al., 2017). Despite the evidence for the validity and reliability of the short versions of the ICU among Chinese community samples, the results may not be generalized to clinical and detained populations. Given that the gravity of juvenile crimes has aggravated in recent years in mainland China, which society has paid more and more attention to, and CU traits are a clinical construct, it is important to expand upon previous findings among different Chinese samples, particularly in detained youths, and test other relevant correlates such as empathy and additional instruments of psychopathic features.

The Current Study

The main purpose of this study was to explore the factor structure of the ICU in a sample of Chinese detained juveniles. Confirmatory factor analyses (CFA) were conducted to compare various factor structures proposed in previous studies. Based on findings from recent studies (Wang et al., 2017b, 2019), we hypothesized that the ICU-11 with the callousness and uncaring dimensions would be the best fit for the data.

The second purpose of this study was to evaluate the psychometric properties of the best-fitted model (ICU-11) including internal consistency and convergent validity. Based on previous research (Wang et al., 2017b, 2019; Deng et al., 2019), it was expected that the ICU-11 would have satisfactory internal consistency while keeping sufficient information from the original 24-item version of the ICU. Additionally, we expected that the ICU-11 scores would correlate positively with alternative instruments of the psychopathic traits (i.e., the Antisocial Process Screening Device – Self-Report Version [APSD-SR] and the Youth Psychopathic Traits Inventory – Short Version [YPI-S]), and the instrument that measures reactive and proactive aggression. Conversely, we expected the scores of the ICU-11 to correlate negatively with empathy (Kimonis et al., 2013). Based on previous findings using indicators of the offending history (Byrd et al., 2013; Pechorro et al., 2017), we expected that the ICU-11 would have correlations with several external criterion variables including the participants' age,

age of incarceration into a juvenile detention center and the duration of incarceration (i.e., difference between current age and first arrest age).

MATERIALS AND METHODS

Participants

The current study included juvenile male participants recruited from the Guangdong Juvenile Detention Center. Excluding participants who had intellectual disability, a total of 613 male participants (*N* = 613, mean age = 17.14, SD = 1.09, range = 14–22) participated voluntarily in the study. Participants were predominantly from nuclear families (*N* = 466, 76.0%), followed by single-parent families (*N* = 135, 22.0%); 79.1% (*N* = 485) came from a multiple-child family. About 64.6% participants (*N* = 396) reported that they had lived with their parents before the age of twelve, followed by grandparents (*N* = 158, 25.8%) and finally, relatives (*N* = 24, 3.9%). With regard to their parents' level of education, 88% of participants' fathers and 92.3% of their mothers were at or below senior secondary school level (similar to Grade 12 in United States). The mean age of participants' first incident of arrest was 15.49 years (SD = 0.87 years). Within the sample, the most common offence committed was robbery (*N* = 411, 67.0%), followed by physical assault (*N* = 70, 11.4%) and sexual assault (*N* = 50, 8.2%)

Procedure

After receiving written informed consent from the detainees' parents or caregivers, the detainees were informed about the aims, content and duration of the study by trained research assistants. They were informed that participation was voluntary, and completion of the study was anonymous. The participants completed the paper-and-pencil self-report survey during their classes, each of which contained 35–40 inmates under the supervision of the research assistants. During the study, participants were allowed to ask for clarification if they did not understand any part of the questionnaire. The study duration was approximately 40 min. This study was approved by the Human Subjects Review Committee at the Guangzhou University. Written informed consent was obtained from all adult participants and from the parents/legal guardians of all non-adult participants.

Measures

Inventory of Callous-Unemotional Traits (ICU; Essau et al., 2006)

The ICU contains 24 items with three factors: callousness (11 items), uncaring (eight items) and unemotional (five items). Each item is rated on a four-point Likert scale, ranging from 1 ("Not at all true") to 4 ("Definitely true"). The higher score indicated a higher endorsement of the item characteristic. The Chinese version of the ICU was created and validated in a sample of Chinese community adults (Wang et al., 2017b), and in that study the Cronbach's αs were 0.80, 0.75, 0.68, and 0.66 for the total score breakdown of callousness, uncaring, and unemotional, respectively.

Antisocial Process Screening Device – Self-Report Version (APSD-SR; Frick and Hare, 2001)

The APSD-SR is a 20-item scale that assesses antisocial behaviors and psychopathic traits in youth. It has three main factors: callous/unemotional (six items), narcissism (seven items) and impulsivity (five items). Each item is rated on a three-point Likert scale from 0 ("Not at all true") to 2 ("Definitely true"). As prior studies with justice-involved youths validated (e.g., Murrie and Cornell, 2002; Pardini et al., 2003), Cronbach's αs ranged from insufficient to acceptable in the current study, 0.71 for the total, 0.44 for the callous-unemotional dimension, 0.61 for the impulsivity dimension, and 0.55 for the narcissism dimension.

Youth Psychopathic Traits Inventory – Short Version (YPI-S; van Baardewijk et al., 2010)

The YPI-S is an 18-item self-report questionnaire that assesses the core psychopathic personality traits (Andershed et al., 2002; Wang et al., 2017a). It consists of three factors: interpersonal (grandiose-manipulative), affective (callous-unemotional), and behavioral (impulsive-irresponsible). Each factor has eight items and each item is scored on a four-point Likert scale ranging from 1 ("Does not apply at all") to 4 ("Applies very well"). Cronbach's αs in the present study were 0.79 for the YPI-S total, 0.76 for the interpersonal scale, and 0.70 for the behavioral scale, but somewhat low (i.e., 0.55) for the affective scale generally consistent with relevant findings (Colins et al., 2012).

Reactive-Proactive Aggression Questionnaire (RPQ; Raine et al., 2006)

The RPQ is a 23-item measure of proactive and reactive aggression in youth and young adults. Reactive aggression is assessed by 11 items, and proactive regression is assessed by 12 items. Each item is rated on a three-point scale from 0 ("Never") to 2 ("Often"). In the present study, Cronbach's αs for the total and factors were 0.94, 0.87, and 0.90, respectively.

Basic Empathy Scale (BES; Jolliffe and Farrington, 2006)

The BES is a 20-item scale that assesses empathy in juveniles. It has two factors: affective empathy (11 items) and cognitive empathy (nine items). Each item is scored on a five-point Likert scale ranging from 1 ("Strongly disagree") to 5 ("Strongly agree"). In the present study, Cronbach's αs for BES total and the two factors (affective and cognitive empathy scales) were 0.74, 0.68, and 0.76, respectively.

Based on standard translation procedures, all above-mentioned measures were adapted and translated into Mandarin Chinese, then back-translated into English by a team led by the second author who is skilled in both Mandarin Chinese and English. Differences in the original and the back-translated versions were discussed and solved by joint agreement of all translators to ensure accuracy.

Data Analysis Strategy

Confirmatory factor analyses were carried out in Mplus 7.4 (Muthén and Muthén, 1998–2015). The factor models examined included the original ICU inter-correlated three-factor model

(M1), the original ICU three-factor bifactor model (M2), the ICU-12 two-factor model (M3), and the ICU-11 two-factor model (M4). The robust weighted least-squares with a mean and variance adjustment (WLSMV) estimator was used to account for the categorical nature of the responses (Flora and Curran, 2004). To assess the model fit, we examined fit indices including chi-square (χ^2), root mean square error of approximation (RMSEA), the Tucker-Lewis index (TLI), and the comparative fit index (CFI). A value of the TLI and CFI at 0.90 or higher and a value of RMSEA at 0.06 or smaller indicate a satisfactory model fit (Kline, 2010).

The internal consistency of the models were assessed by computing Cronbach's α values as well as the mean inter-item correlations (MIC), a more straightforward indicator regardless of the length of a scale. Conventional guidelines suggest that the Cronbach's α values $\geq$ 0.70 indicate acceptable internal consistency (Barker et al., 1994) and a MIC value between 0.15 and 0.50 indicates satisfactory internal consistency (Clark and Watson, 1995). To provide a more rigorous evaluation of the internal reliability of the ICU versions based on CFA models, we also investigated the composite reliability of the measurement properties of the scale. A value greater than 0.60 is generally considered acceptable (Bagozzi and Yi, 1988; Diamantopoulos and Siguaw, 2000). The convergent and discriminant validity evaluated via Pearson's correlations were between the ICU scores and criterion variables (e.g., APSD-SR, YPI-S, RPQ and BES). We analyzed the internal consistency and correlations of the models using the SPSS program (IBM, SPSS version 19, 2010). Finally, the method proposed by Dunn and Clark (1969) was used (see Steiger, 1980 for more details)[1] to determine whether the strength of the correlations with criterion measures differed between the original ICU and the best-fit model of ICU.

RESULTS

Table 1 reports descriptive statistics including means, standard deviations, number of items as well as Cronbach's α values and MICs about all variables in the currents study.

Confirmatory Factor Analysis

Table 2 shows the fit indices of competitive models used in the current study. Fit indices showed an unacceptable fit for the inter-correlated three-factor model (M1; χ^2 = 1901.46, df = 249, CFI = 0.71, TLI = 0.68, RMSEA = 0.10) and for the original three-factor bifactor (M2; χ^2 = 1930.16, df = 228, CFI = 0.70, TLI = 0.64, RMSEA = 0.11). The two-factor model of the ICU-12 had significantly better fit than the M1 or M2, but the fit indices were still unsatisfactory (CFI < 0.90, TLI < 0.90, RMSEA > 0.80). Moreover, Item Six had the lowest loading (λ = 0.26, see **Table 3**). The two-factor model (ICU-11) that excluded Item Six had an excellent fit (χ^2 = 149.77, df = 43; CFI = 0.95, TLI = 0.94, RMSEA = 0.06).

With regards to the internal consistency, the Cronbach's αs (MICs) for the ICU-11 total score, the callousness factor and uncaring factor were 0.75 (MIC = 0.22), 0.75 (MIC = 0.34), and 0.73 (MIC = 0.35), respectively. Furthermore, the results showed that all factor scores of the ICU-11 were measured with satisfactory composite reliability (total score, ρ_c = 0.90;

TABLE 1 | Descriptive statistics and reliability estimates for all variables.

	Mean	SD	MIC	α	N
ICU-24					
Unemotional	13.36	2.48	0.13	0.41	5
Callousness	18.64	4.99	0.25	0.77	11
Uncaring	17.78	4.70	0.35	0.81	8
Total	49.78	8.28	0.13	0.77	24
ICU-12					
Callousness	11.26	3.59	0.29	0.73	7
Uncaring	10.95	3.19	0.35	0.73	5
Total	22.20	5.20	0.19	0.73	12
ICU-11					
Callousness	9.02	3.19	0.34	0.75	6
Uncaring	10.95	3.19	0.35	0.73	5
Total	19.96	5.02	0.22	0.75	11
APSD-SR					
Impulsivity	3.35	2.24	0.24	0.61	5
CU	3.69	2.02	0.13	0.44	6
Narcissism	3.47	2.28	0.16	0.55	7
Total	11.71	5.26	0.11	0.71	20
YPI-S					
Behavioral factor	11.31	3.69	0.34	0.76	6
Affective factor	10.70	3.04	0.17	0.55	6
Interpersonal factor	8.80	2.75	0.30	0.70	6
Total	30.71	7.14	0.18	0.79	18
RPQ					
Reactive	6.84	4.63	0.39	0.87	11
Proactive	5.03	5.00	0.43	0.90	12
Total	11.84	9.13	0.39	0.94	23
BES					
Affective	34.69	6.04	0.16	0.69	11
Cognitive	33.50	5.32	0.26	0.76	9
Total	68.26	8.79	0.12	0.74	20

ICU-24, Inventory of Callous and Unemotional Traits; ICU-12, Inventory of Callous and Unemotional Traits – 12 items, short version; ICU-11, Inventory of Callous and Unemotional Traits – 11 items, short version; APSD-SR, Antisocial Process Screening Device – self-report version; CU, Callous-Unemotional Traits; YPI-S, Youth Psychopathic Traits Inventory – short version; RPQ, Reactive-Proactive Aggression Questionnaire; BES, Basic Empathy Scale; SD, standard deviation; MIC, mean inter-item correlation; N, number of items.

TABLE 2 | Goodness-of-fit indices for the different models of ICU.

	WLSMVχ^2	df	RMSEA (90% CI)	CFI	TLI
M1	1901.46***	249	0.10 [0.10, 0.11]	0.71	0.68
M2	1930.16***	228	0.11 [0.11, 0.12]	0.70	0.64
M3	302.34***	53	0.09 [0.08, 0.10]	0.89	0.87
M4	149.77***	43	0.06 [0.05, 0.08]	0.95	0.94

*M1, inter-correlated three-factor model; M2, original three-factor bifactor model; M3, ICU-12; M4, ICU-11; WLSMV, weighted least squares with mean and variance; df, degrees of freedom; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; CFI, Comparative Fit Index; TLI, Tucker–Lewis Index. ***p < 0.001.*

[1] Using a spreadsheet that was developed by DeCoster and Iselin (2005) and can be retrieved at: http://stat-help.com/spreadsheets.html

TABLE 3 | Factor loadings for the relatively good fit two-factor model for ICU-12 and ICU-11.

Items	Callousness	Uncaring
(4) I do not care who I hurt to get what I want	0.72/0.72	
(6) I do not show my emotions to others	0.26	
(9) I do not care if I get into trouble	0.75/0.74	
(11) I do not care about doing things well	0.60/0.59	
(12) I seem very cold and uncaring to others	0.65/0.62	
(18) I do not feel remorseful when I do something wrong	0.61/0.61	
(21) The feelings of others are unimportant to me	0.78/0.79	
(5) I feel bad or guilty when I do something wrong (R)		0.73/0.74
(8) I am concerned about the feelings of others (R)		0.64/0.64
(16) I apologize (say "I am sorry") to persons I hurt (R)		0.74/0.73
(17) I try not to hurt others' feelings (R)		0.58/0.58
(24) I do things to make others feel good (R)		0.56/0.56

ICU-12, Inventory of Callous-Unemotional Traits – 12 items, short version; ICU-11, Inventory of Callous-Unemotional Traits – 11 items, short version; (R), negatively worded items reverse-scored prior to analysis; factor loadings of ICU-11 are presented after the slash; all factor loadings are significant at a level of 0.001.

callousness, ρ_c = 0.84; uncaring, ρ_c = 0.79). The correlation between the two factors was.24 ($p < 0.001$) at the observed level and 0.21 ($p < 0.001$) at the latent variable level, indicating a relatively weak intercorrelation.

Convergent and Criterion Validity

Table 4 shows Pearson's correlations between the ICU-11 and external criterion measures. As expected, there were significantly positive correlations between the ICU-11 factors and APSD-SR factors. The ICU-11 uncaring factor had a strong correlation with the APSD-SR callous/unemotional factor ($r = 0.50$, $p < 0.001$). The ICU-11 callousness factor was strongly correlated with the APSD-SR impulsiveness factor as well as the APSD-SR total (r = 0.50 and 0.53, ps < 0.001, respectively). The ICU-11 callousness factor showed significantly positive correlations with the YPI-S total scores and factors (rs = 0.45–0.67, ps < 0.001). On the other hand, the ICU-11 uncaring factor had weak correlations with the YPI-S behavioral factor and YPI-S total scores ($r = 0.22$, $p < 0.001$, and 0.11, $p < 0.05$, respectively), and was not significantly correlated with the YPI-S affective ($r = -0.02$, $p > 0.05$) or interpersonal factors ($r = -0.04$, $p > 0.05$).

The ICU-11 total score and the ICU-11 callousness scale were moderately and positively correlated with two kinds of aggression assessed by RPQ (see **Table 4**). On the other hand, the ICU-11 uncaring scale showed weak associations with aggression (rs < 0.30). The ICU-11 total also had a significant negative correlation with empathy as measured by the BES (total BES: $r = -0.51$, $p < 0.001$; affective factor: $r = -0.35$, $p < 0.001$; cognitive factor: $r = -0.45$, $p < 0.001$). The ICU-11 uncaring factor had stronger relationships with the BES and its factors (rs = −0.32 to −0.45, ps < 0.001) than the ICU-11 callousness factor did (r = −0.24 to -0.35, ps < 0.001).

Correlations between the original ICU total and factor scores and external variables were similar to those for the ICU-11 (see **Table 4**). The unemotional factor of the original ICU demonstrated weaker or no associations at all with the external variables, whereas it showed robustly stronger associations with scores for reactive aggression, the YPI-S behavioral factor, proactive aggression and the APSD-SR narcissism factor.

Table 4 also presents the correlations between the ICU-11 and other variables (e.g., age, age of incarceration into a juvenile detention center). The ICU-11 and subscale scores were negatively correlated with age, but positively correlated with the age of incarceration. To explore this further, we inspected the correlations between the ICU-11 and the duration of incarceration (i.e., difference between current age and first arrest age). There was a significant negative correlation between the ICU-11 and the duration of incarceration, suggesting that participants with a longer stay at the center reported lower ICU scores. The original ICU were as and the ICU-11 had similar correlations with those variables.

Next, we compared the ICU-11 and the original ICU in terms of their correlations with the external criterion variables. Z values ($p < 0.01$, two-tailed for significance) were calculated based on Dunn and Clark (1969) method (see **Table 4**). For most variables, the ICU-11 total showed stronger correlations to the external criterion than the ICU-24 did.

DISCUSSION

The present study is the first study that investigated the factor structure and psychometric properties of the ICU in Chinese detained youth samples. Consistent with previous studies using samples of Chinese community adults (Wang et al., 2017b) and children (Wang et al., 2019), the three-factor model of the original ICU was not replicated in the present study, but the ICU-11 with a two-factor model was found to have the best fit for the data. The reliability coefficients of the ICU-11 and its factors were also more satisfying than those of the original ICU. Finally, the convergent validity of the ICU was demonstrated by significant correlations between the ICU-11 and a range of criteria variables.

Previous studies of the ICU using Western samples found that the three-factor bifactor model received the most support in adolescents (Kimonis et al., 2008; Pihet et al., 2015). However, the bifactor model could not be replicated in the current study as well as it could with other Chinese samples (Wang et al., 2017b). The poor fit was mainly attributed to the low factor loading of items on the unemotional factor. Additionally, the unemotional factor of the original ICU-24 showed substantially low Cronbach's α value and poor validity, which was in line with previous studies (Kimonis et al., 2008; Byrd et al., 2013; Wang et al., 2017b; Deng et al., 2019). Despite the unemotional factor showing high association with empathy and modest association with proactive

TABLE 4 | Pearson correlations of ICU-11, ICU-24, and their factors with relevant external variables.

	ICU-11			ICU-24				*Z*
APSD-SR	**Uncaring**	**Callousness**	**Total**	**Unemotional**	**Uncaring**	**Callousness**	**Total**	**ICU-24 Total vs. ICU-11 Total**
Impulsivity	0.16***	0.50***	0.42***	−0.07	0.18***	0.55***	0.42***	0.00
CU	0.50***	0.29***	0.50***	−0.01	0.51***	0.37***	0.52***	1.37
Narcissism	0.15***	0.38***	0.33***	−0.12**	0.13**	0.42***	0.29***	−2.46*
Total	0.35***	0.53***	0.56***	−0.10*	0.36***	0.61***	0.55***	−0.71
YPI-S scores								
Behavioral factor	0.22***	0.58***	0.50***	−0.19***	0.23***	0.64***	0.46***	−2.68**
Affective factor	−0.02	0.46***	0.29***	−0.04	−0.06	0.42***	0.24***	−3.03**
Interpersonal factor	−0.04	0.45***	0.26***	−0.08	−0.08	0.40***	0.17***	−5.39***
Total	0.11*	0.67***	0.49***	−0.14**	0.08	0.66***	0.41***	−5.28***
RPQ scores								
Reactive	0.21***	0.47***	0.43***	−0.21***	0.21***	0.52***	0.37***	−3.84***
Proactive	0.21***	0.50***	0.44***	−0.14**	0.20***	0.52***	0.39***	−3.22**
Total	0.23***	0.51***	0.46***	−0.18***	0.22***	0.55***	0.41***	−3.26**
BES scores								
Affective	−0.32***	−0.24***	−0.35***	−0.09*	−0.25***	−0.24***	−0.32***	1.86
Cognitive	−0.39***	−0.30***	−0.45***	0.08	−0.38***	−0.35***	−0.42***	1.95
Total	−0.45***	−0.35***	−0.51***	−0.02	−0.40***	−0.38***	−0.47***	2.70**
Age	−0.06	−0.11**	−0.12**	−0.05	−0.08	−0.12**	−0.14**	−1.17
AIJDC	0.08	0.12**	0.13**	−0.01	0.10*	0.11**	0.13**	0.00
DI	−0.14***	−0.21***	−0.23***	−0.05	−0.16***	−0.22***	−0.25***	−1.20

ICU-24, Inventory of Callous and Unemotional Traits; ICU-11, Inventory of Callous and Unemotional Traits – 11 items, short version; APSD-SR, Antisocial Process Screening Device – self-report version; CU, Callous-Unemotional Traits; YPI-S, Youth Psychopathic Traits Inventory – short version; RPQ, Reactive-Proactive Aggression Questionnaire; BES, Basic Empathy Scale; AIJDC, Age of incarceration into a Juvenile Detention Center; DI, duration of incarceration. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

aggression across over ten studies (Cardinale and Marsh, 2017), these findings were hardly replicated in this Chinese detained juvenile sample thus to some extent indicated the unemotional were not a stable indicator of the construct of CU traits and needed further validation.

These results have reinforced the idea that the original unemotional factor of the ICU might not be a reliable construct in detained youth, at least when using the self- or other-report versions of the ICU. A major reason for this is considered to be that the affective deficits lack accurate descriptions, and that most items looking at the unemotional factor refer to the outward expression of emotions rather than the experience of them, both of which result in poor internal consistency in the unemotional factor (Cardinale and Marsh, 2017). The features of unemotional trait are mostly negative, which are more difficult to detect for both the subjects and the observers. Subjects may not be aware of the absence of emotion, while observers may mistake the symptoms as the subject being shy or introverted. Another factor is that the expressions of "unemotional" characteristics could also be contributed to by other constructs, such as social expectations or problematic emotional expressions (such as those by autistic children). Social expectations vary greatly across cultures and, thus, can negatively influence the multigroup measurement invariance across the original English samples, as well as subsequent samples from other cultural groups. All these issues could result in lower reliability of the unemotional factor.

With regards to problematic emotional expressions, previous studies have consistently found negative correlations of the unemotional factor with aggression assessments (Wang et al., 2017b). Subjects with abnormal emotional regulation and expression may externalize emotions such as anger, demonstrating aggressive behaviors. Taken together, the items of the unemotional factor may be tapping into a construct departing from CU. Further research into the unemotional factor is warranted.

The shortened ICU-12 that excluded most items from the unemotional factor achieved a better fit than the original ICU factor structures, with the exception of Item 6, which had a low factor loading. This was consistent with previous studies (Colins et al., 2016; Wang et al., 2017b, 2019). After removing Item 6, the ICU-11 had the best fit for the current data.

The analysis of the internal consistency of the ICU-11 revealed mostly good to extremely good values, with most values exceeding both the recommended minimum Cronbach's α of 0.70 and the recommended minimum composite reliability of 0.60, as well as the MICs in a favorable range (>0.19). The Cronbach's α values of both the ICU-12 and the ICU-11 uncaring factors in the present study were greater than in previous findings (Wang et al., 2017b, 2019). The greater factor reliability could be due to the fact that the sample for this study had an older average age than studies where the sample consisted of children. Adolescent subjects in the present study might have had better reading comprehension than those under the age of 12 years (Soto et al., 2008; Deng et al., 2019). In addition, the ICU was developed based on a clinical sample, thus could be more precise when measuring CU traits among subjects who were on the high end of the latent traits. And, in comparison to community samples, the detention environment helped to guarantee the standardization of the testing process, which may have offered more consistent responses to the ICU items. Furthermore, it was worth mentioning that the α values for ICU scores in clinical samples had been proven to be more variable than in non-clinical samples (Deng et al., 2019). More evidence for internal consistency of ICU-11 in Chinese clinical samples is needed in the future.

With regards to external validity, the ICU-11 demonstrated the expected correlations with the criterion variables (i.e., APSD-SR, YPI-S, and RPQ), and the pattern of correlations were similar to those of the original ICU.

As reported by previous findings of a meta-analytic review (Cardinale and Marsh, 2017), strong associations were found between psychopathy and the total ICU-11, callousness factor and uncaring factor, and the callousness factor compared with the uncaring factor displayed stronger associations with measures of psychopathy in detained samples. Specifically, the directions and magnitudes of the correlations between the ICU and the YPI-S were comparable with those reported in previous studies (Roose et al., 2010; Pihet et al., 2015). Most correlations found between the ICU-11 scales and APSD-SR scales were higher than those reported in Wang et al. (2017b), which reflects the different demographics of the two samples. Wang et al. (2017b) used a community sample, in which the manifest of antisocial personality had a limited range.

Meanwhile, consistent with previous studies, the aggression factor showed a stronger correlation with callousness than with the uncaring factor. Kimonis et al. (2008) suggested that this could be due to the fact that callousness has a greater comorbidity with aggression, whereas uncaring was expressed through their offences committed. The ICU-11 also demonstrated expected negative associations with empathy when assessed by the BES (e.g., Kimonis et al., 2008; Roose et al., 2010). Dolan and Fullam (2006) suggested that the temperamental fearlessness featured in CU traits can result in a decrease in the arousal of the autonomic nervous system. This in turns leads to difficulties in recognizing others' emotional distress among individuals who rank high in psychopathy measurements. The uncaring factor also had stronger correlations with the BES than the callousness, suggesting that the uncaring is a major component in one's inability to recognize others' emotions. Similar findings were also reported by Pechorro et al. (2016a, 2017).

We also evaluated how the CU traits were related to subjects' age, age of incarceration, and the duration of incarceration. Inconsistent with previous findings (Byrd et al., 2013; Pechorro et al., 2017), we found that the CU traits had moderately negative associations with participants' age and the duration of incarceration. This suggested that older participants might be better at identifying and reporting emotion. In addition, Asscher et al. (2011) indicated that individual age when assessing psychopathy played a moderating role in the associations between psychopathy and delinquency. Notably, during the course of childhood to adolescence, individuals with psychopathic traits likely have learned to conceal their cognitive empathy deficits or the relevant empathy skills may have improved (Dadds et al., 2009). Thus, the strength of association between psychopathy and delinquency diminished with increasing age (Asscher et al., 2011). Overall, the incarceration confinement and education seemed to have a positive effect on transforming the pathological personality of the juvenile offenders.

Summarizing, prior findings have emphasized the importance of CU traits which appear to mirror several related aspects about affective and interpersonal functioning (Lynam et al., 2005). CU traits also provide evidence to designate and understand severely antisocial youths, especially the adolescent offenders who had great risk in subsequent violent offenses throughout a 2-year period after releasing from incarceration (Vincent et al., 2003). Currently in China, market reforms have promoted the social transition, meanwhile, the crime rate of juveniles has assumed the trend of escalation and criminal nature of the case has become more and more serious. Assessment of CU traits with the ICU particularly the shortened ICU-11 thus remains a significant research focus with crucial clinical implications in Chinese juvenile offenders. Specifically, extant findings may allow psychological staff to tap Chinese detained boys the existence of the common factor, analyze the causes of crime or delinquency and thus take appropriate measures to improve the system of current criminal penalty.

Limitations

Several limitations must be acknowledged. First, the current sample was made up only of males, making it unclear how the results can be generalized toward female detention populations. Pechorro et al. (2017) found manifestations of generalized problem conducts in female juveniles with CU traits might depend on the criminal justice system. Future study should look at female populations and examine potential gender differences regarding the validity and reliability of the ICU. Second, all measures were based on self-reporting and the current study did not explore the detailed offending history of the detained boys, which easily demonstrated method variance and might inflate relations among study variables. Future research should consider the inclusion of multiple methods of data gathering, such as interviews, multiple-informant formats, such as caregiver- or caseworker-reported, and include more delinquent details from case records. Third, the current study had a cross-sectional design, which restricted the conclusions on the predictive utility of ICU traits, as well as any causal inferences. Future longitudinal studies should be conducted that evaluate correlations over time. Finally, future research also should investigate the relationships between the ICU-11 and variables such as delinquent histories, conduct disorder, age of first contact with the law, and the severity of the crime.

CONCLUSION

The current study is the first study to explore the factor structure and construct validity of the ICU in a large Chinese male juvenile offender sample. Consistent with previous studies looking at Chinese samples (Wang et al., 2017b, 2019), CFA analyses indicated that the ICU-11 with two factors had the best model fit. Both the total and two factors' scores showed acceptable internal consistency. The results also demonstrated promising convergent validity of the ICU-11. Overall, the current study's findings suggest that the ICU-11 holds promise as an informative alternative for the original ICU form, particularly in detained Chinese male youths.

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

After receiving written informed consent from the detainees' parents or caregivers, the detainees were informed about the aims, content, and duration of the study by trained research assistants. The study duration was approximately 40 min. This study was approved by the Human Subjects Review Committee at Guangzhou.

AUTHOR CONTRIBUTIONS

XZ, YS, CZ, JL, and WY made substantial contribution to the analysis and interpretation of the data, drafted the manuscript, provided the final approval for the manuscript, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. M-CW and YG made substantial contributions to the conception and the design of the study, drafted the manuscript, provided final approval for the manuscript, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 31800945 and 31400904) and Guangzhou University's 2017 training program for young top-notch personnels (BJ201715).

REFERENCES

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Arlington, VA: American Psychological Association.

Andershed, H., Kerr, M., Stattin, H., and Levander, S. (2002). "Psychopathic traits in non–referred youths: initial test of a new assessment tool," in *Psychopaths: Current International Perspectives*, eds E. Blaauw, J. M. Philippa, K. C. M. P. Ferenshild, and B. Van Lodesteijn (The Hague: Elsevier), 131–158.

Asscher, J. J., Van Vugt, E. S., Stams, G. J. J. M., Deković, M., Eichelsheim, V. I., and Yousfi, S. (2011). The relationship between juvenile psychopathic traits, delinquency and (violent) recidivism: a meta-analysis. *J. Child Psychol. Psychiatry* 52, 1134–1143. doi: 10.1111/j.1469-7610.2011.02412.x

Bagozzi, R. P., and Yi, Y. (1988). On the evaluation of structural equation models. *J. Acad. Mark. Sci.* 16, 74–94. doi: 10.1007/bf02723327

Barker, C., Pistran, N., and Elliot, R. (1994). *Research Methods in Clinical and Counselling Psychology*. Chichester: Wiley.

Byrd, A. L., Kahn, R. E., and Pardini, D. A. (2013). A validation of the inventory of callous-unemotional traits in a community sample of young adult males. *J. Psychopathol. Behav. Assess.* 35, 20–34. doi: 10.1007/s10862-012-9315-4

Cardinale, E. M., and Marsh, A. A. (2017). The reliability and validity of the inventory of callous unemotional traits: a meta-analytic review. *Assessment* doi: 10.1177/1073191117747392 [Epub ahead of print].

Ciucci, E., Baroncelli, A., Franchi, M., Golmaryami, F. N., and Frick, P. J. (2014). The association between callous-unemotional traits and behavioral and academic adjustment in children: further validation of the inventory of callous-unemotional traits. *J. Psychopathol. Behav. Assess.* 36, 189–200. doi: 10.1007/s10862-013-9384-z

Clark, L. A., and Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychol. Assess.* 7, 309–319. doi: 10.1037/1040-3590.7.3.309

Colins, O., Noom, M., and Vanderplasschen, W. (2012). Youth psychopathic traits inventory - short version: a further test of the internal consistency and criterion validity. *J. Psychopathol. Behav. Assess.* 34, 476–486. doi: 10.1007/s10862-012-9299-0

Colins, O. F., Andershed, H., Hawes, S. W., Bijttebier, P., and Pardini, D. A. (2016). Psychometric properties of the original and short form of the inventory of callous-unemotional traits in detained female adolescents. *Child Psychiatry Hum. Dev.* 47, 679–690. doi: 10.1007/s10578-015-0601-8

Corrado, R. R., DeLisi, M., Hart, S. D., and McCuish, E. C. (2015). Can the causal mechanisms underlying chronic, serious, and violent offending trajectories be elucidated using the psychopathy construct? *J. Crim. Justice* 43, 251–261. doi: 10.1016/j.jcrimjus.2015.04.006

Dadds, M. R., Hawes, D. J., Frost, A. D., Vassallo, S., Bunn, P., Hunter, K., et al. (2009). Learning to 'talk the talk': the relationship of psychopathic traits to deficits in empathy across childhood. *J. Child Psychol. Psychiatry* 50, 599–606. doi: 10.1111/j.1469-7610.2008.02058.x

DeCoster, J., and Iselin, A.-M. (2005). *Comparing Correlation Coefficients [Spreadsheet]*. Available at: http://stat-help.com/spreadsheets.html (accessed December 31, 2018).

DeLisi, M. (2016). *Psychopathy as Unified Theory of Crime*. Basingstoke: Palgrave Macmillan.

DeLisi, M., Fox, B. H., Fully, M., and Vaughn, M. G. (2018). The effects of temperament, psychopathy, and childhood trauma among delinquent youth: a test of Delisi and Vaughn's temperament-based theory of crime. *Int. J. Law Psychiatry* 57, 53–60. doi: 10.1016/j.ijlp.2018.01.006

DeLisi, M., and Vaughn, M. (2015). Ingredients for criminality require genes, temperament, and psychopathic personality. *J. Crim. Justice* 43, 290–294. doi: 10.1016/j.jcrimjus.2015.05.005

DeLisi, M., Vaughn, M., Beaver, K. M., Wexler, J., Barth, A. E., and Fletcher, J. M. (2010). Fledgling psychopathy in the classroom: ADHD subtypes, psychopathy, and reading comprehension in a community sample of adolescents. *Youth Violence Juv. Justice* 9, 43–58. doi: 10.1177/1541204010371932

Deng, J. X., Wang, M.-C., Zhang, X., Shou, Y., Gao, Y., and Luo, J. (2019). The inventory of callous unemotional traits: a reliability generalization meta-analysis. *Psychol. Assess.* 31, 765–780. doi: 10.1037/pas0000698

Diamantopoulos, A., and Siguaw, J. A. (2000). *Introducing LISREL*. London: Sage Publications.

Dolan, M., and Fullam, R. (2006). Face affect recognition deficits in personality-disordered offenders: association with psychopathy. *Psychol. Med.* 36, 1563–1569. doi: 10.1017/S0033291706008634

Dunn, O. J., and Clark, V. (1969). Correlation coefficients measured on the same individuals. *J. Am. Stat. Assoc.* 64, 366–377. doi: 10.1080/01621459.1969.10500981

Essau, C. A., Sasagawa, S., and Frick, P. J. (2006). Callous-unemotional traits in community sample of adolescents. *Assessment* 13, 454–469. doi: 10.1177/1073191106287354

Fisher, L., and Blair, R. J. R. (1998). Cognitive impairment and its relationship to psychopathic tendencies in children with emotional and behavioral difficulties. *J. Abnorm. Child Psychol.* 26, 511–519. doi: 10.1023/A:1022655919743

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods* 9, 466–491. doi: 10.1037/1082-989X.9.4.466

Frick, P. J. (2004). *Inventory of Callous-Unemotional Traits*. New Orleans, LA: University of New Orleans.

Frick, P. J. (2009). Extending the construct of psychopathy to youth: implications for understanding, diagnosing, and treating antisocial children and adolescents. *Can. J. Psychiatry* 54, 803–812. doi: 10.1192/bjp.195.6.562

Frick, P. J., and Hare, R. (2001). *The Antisocial Process Screening Device (APSD): Technical Manual*. Toronto, ON: Multi-health Systems.

Frick, P. J., Lilienfeld, S. O., Ellis, M., Loney, B., and Silverthorn, P. (1999). The association between anxiety and psychopathic traits dimensions in children. *J. Abnorm. Child Psychol.* 27, 383–392. doi: 10.1023/A:1021928018403

Frick, P. J., and Moffitt, T. E. (2010). *A Proposal to the DSM–V Childhood Disorders and the ADHD and Disruptive Behavior Disorders Work Groups to Include a Specifier to the Diagnosis of Conduct Disorder Based on the Presence of Callous-Unemotional Traits*. Washington, DC: American Psychiatric Association.

Hawes, S. W., Byrd, A. L., Henderson, C. E., Gazda, R. L., Burke, J. D., Loeber, R., et al. (2014). Refining the parent-reported inventory of Callous-unemotional traits in boys with conduct problems. *Psychol. Assess.* 26, 256–266. doi: 10.1037/a0034718

Jolliffe, D., and Farrington, D. P. (2006). Development and validation of the basic empathy scale. *J. Adolesc.* 29, 589–611. doi: 10.1016/j.adolescence.2005.08.010

Kahn, R. E., Frick, J. P., Youngstrom, E., Findling, R. L., and Youngstrom, J. K. (2012). The effects of including a Callous-Unemotional specifier for the diagnosis of conduct disorder. *J. Child Psychol. Psychiatry* 53, 271–282. doi: 10.1111/j.1469-7610.2011.02463.x

Kimonis, E. R., Branch, J., Hagman, B., Graham, N., and Miller, C. (2013). The psychometric properties of the Inventory of Callous-Unemotional traits in an undergraduate sample. *Psychol. Assess.* 25, 84–93. doi: 10.1037/a0029024

Kimonis, E. R., Frick, P. J., Skeem, J. L., Marsee, M. A., Cruise, K., Munoz, L. C., et al. (2008). Assessing callous-unemotional traits in adolescent offenders: validation of the inventory of callous-unemotional traits. *Int. J. Law Psychiatry* 31, 241–252. doi: 10.1016/j.ijlp.2008.04.002

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling*, 3rd Edn. New York, NY: Guilford Press.

Lynam, D. R., Caspi, A., Moffitt, T. E., Raine, A., Loeber, R., and Stouthamer-Loeber, M. (2005). Adolescent psychopathy and the big five: results from two samples. *J. Abnorm. Child Psychol.* 33, 431–443. doi: 10.1007/s10648-005-5724-0

Murrie, D., and Cornell, D. (2002). Psychopathy screening of incarcerated juveniles: a comparison of measures. *Psychol. Assess.* 14, 390–396. doi: 10.1037/1040-3590.14.4.390

Muthén, L. K., and Muthén, B. O. (1998–2015). *Mplus User's Guide*, 7th Edn. Los Angeles, CA: Muthén & Muthén. doi: 10.1037//1040-3590.14.4.390

Paiva-Salisbury, M. L., Gill, A. D., and Stickle, T. R. (2017). Isolating trait and method variance in the measurement of callous and unemotional traits. *Assessment* 24, 763–771. doi: 10.1177/1073191115624546

Pardini, D., Lochman, J., and Frick, P. (2003). Callous/Unemotional traits and social cognitive processes in adjudicated youth. *J. Am. Acad. Child Adolesc. Psychiatry* 42, 364–371. doi: 10.1097/00004583-200303000-00018

Pechorro, P., Hawes, S. W., Gonçalves, R. A., and Ray, J. V. (2016a). Psychometric properties of the inventory of Callous-unemotional traits short version (ICU-12) among detained female juvenile offenders and community youths. *Psychol. Crime Law* 23, 221–239. doi: 10.1080/1068316x.2016.1239724

Pechorro, P., Ray, J. V., Barroso, R., Maroco, J., and Abrunhosa Goncalves, R. (2016b). Validation of the inventory of Callous-unemotional traits among a Portuguese sample of detained juvenile offenders. *Int. J. Offender Ther. Comp. Criminol.* 60, 349–365. doi: 10.1177/0306624X14551256

Pechorro, P., Ray, J. V., Gonçalves, R. A., and Jesus, S. N. (2017). The inventory of Callous–unemotional traits: psychometric properties among referred and non-referred Portuguese female juveniles. *Int. J. Law Psychiatry* 54, 67–75. doi: 10.1016/j.ijlp.2017.05.002

Pihet, S., Etter, S., Schmid, M., and Kimonis, E. R. (2015). Assessing Callous-unemotional traits in adolescents: validity of the inventory of Callous-unemotional traits across gender, age, and community/institutionalized status. *J. Psychopathol. Behav. Assess.* 37, 407–421. doi: 10.1007/s10862-014-9472-8

Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., et al. (2006). The reactive–proactive aggression questionnaire: differential correlates of reactive and proactive aggression in adolescent boys. *Aggress. Behav.* 32, 159–171. doi: 10.1002/ab.20115

Roose, A., Bijttebier, P., Decoene, S., Claes, L., and Frick, P. J. (2010). Assessing the affective features of psychopathy in adolescence: a further validation of the inventory of Callous and unemotional traits. *Assessment* 17, 44–57. doi: 10.1177/1073191109344153

Sherretts, N., Boduszek, D., Debowska, A., and Willmott, D. (2017). Comparison of murderers with recidivists and first time incarcerated offenders from U.S. prisons on psychopathy and identity as a criminal: an exploratory analysis. *J. Crim. Justice* 51, 89–92. doi: 10.1016/j.jcrimjus.2017.03.002

Soto, C. J., John, O. P., Gosling, S. D., and Potter, J. (2008). The developmental psychometrics of big five self-reports: acquiescence, factor structure, coherence, and differentiation from ages 10 to 20. *J. Pers. Soc. Psychol.* 94, 718–737. doi: 10.1037/0022-3514.94.4.718

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychol. Bull.* 87, 245–251. doi: 10.1037/0033-2909.87.2.245

van Baardewijk, Y., Andershed, H., Stegge, H., Nilsson, K. W., Scholte, E., and Vermeiren, R. (2010). Development and tests of short versions of the youth psychopathic traits inventory and the youth psychopathic traits inventory-child version. *Eur. J. Psychol. Assess.* 26, 122–128. doi: 10.1027/1015-5759/a000017

Vaughn, M. G., and DeLisi, M. (2008). Were Wolfgang's chronic offenders psychopaths? On the convergent validity between psychopathy and career criminality. *J. Crim. Justice* 36, 33–42. doi: 10.1016/j.jcrimjus.2007.12.008

Vincent, G. M., Vitacco, M. J., Grisso, T., and Corrado, R. R. (2003). Subtypes of adolescent offenders: affective traits and antisocial behavior patterns. *Behav. Sci. Law* 21, 695–712. doi: 10.1002/bsl.556

Wang, M.-C., Colins, O. F., Deng, Q., Andershed, H., Deng, J., and Ye, H. (2017a). Psychometric properties of the original and shortened version of the youth psychopathic traits inventory among Chinese adolescents. *J. Psychopathol. Behav. Assess.* 39, 620–634. doi: 10.1007/s10862-017-9619-5

Wang, M.-C., Gao, Y., Deng, J., Lai, H., Deng, Q., and Armour, C. (2017b). The factor structure and construct validity of the inventory of callous-unemotional traits in Chinese undergraduate students. *PLoS One* 12:e0189003. doi: 10.1371/journal.pone.0189003

Wang, M.-C., Shou, Y., Liang, J., Lai, H., Zeng, H., Chen, L., et al. (2019). Further validation of the inventory of Callous-unemotional traits: cross-informants invariance and longitudinal invariance. *Assessment* doi: 10.1177/1073191119845052 [Epub ahead of print].

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

BRIEF RESEARCH REPORT
published: 24 September 2019
doi: 10.3389/fpsyg.2019.02134

Flexibility in Existential Beliefs and Worldview: Testing Measurement Invariance and Factorial Structure of the Existential Quest Scale in an Italian Sample of Adults

Marco Rizzo[1], *Silvia Testa*[2]*, *Silvia Gattino*[1] *and Anna Miglietta*[1]

[1] Department of Psychology, University of Turin, Turin, Italy, [2] Department of Human and Social Sciences, University of Aosta Valley, Aosta, Italy

The aim of the present study was to assess the psychometric properties of the Existential Quest (EQ) Scale, a nine-items instrument developed to assess openness to changing one's own convictions concerning existential issues. We developed the Italian version of the scale and examined factorial structure, internal consistency, discriminant validity, and measurement invariance across gender and age groups. A total of 291 Italian adults were recruited, and they completed a self-report questionnaire comprising measures of authoritarianism, cognitive closure, well-being, and religiousness, alongside the EQ. Confirmatory factor analysis showed that the original one-factor structure was replicated in this study, except for one-item that was removed from the subsequent analyses. Both the internal consistency of the eight-item scale as assessed by Cronbach's α and discriminant validity were in line with those of the original study. However, McDonald's reliability coefficient were quite low, and further researches employing repeated measures are needed in order to comprehend the contribution of the random error and that of the item specificity in lowering McDonald's coefficient. Finally, evidence of full measurement invariance across gender and partial measurement invariance across age was obtained. Overall, these findings suggest that the Italian version of the EQ is a promising tool for assessing flexibility about existential issues.

Keywords: Existential Quest Scale, existential beliefs, psychometric properties, factorial structure, measurement invariance

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Caterina Primi,
University of Florence, Italy
Cristina Senín-Calderón,
University of Cádiz, Spain

****Correspondence:***
Silvia Testa
s.testa@univda.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *18 May 2019*
Accepted: *03 September 2019*
Published: *24 September 2019*

Citation:
Rizzo M, Testa S, Gattino S and Miglietta A (2019) Flexibility in Existential Beliefs and Worldview: Testing Measurement Invariance and Factorial Structure of the Existential Quest Scale in an Italian Sample of Adults. Front. Psychol. 10:2134. doi: 10.3389/fpsyg.2019.02134

INTRODUCTION

Addressing the fundamental questions of existence – such as the origin and finality of the world, the meaning of life and death, or the existence of transcendence – is a universal human experience that crosses cultures, historical periods, religions, and ideologies, and may be important for optimal individual functioning (Allan and Shearer, 2012; Sullivan, 2013). Indeed, the exploration of existential issues represents a valuable dimension in the promotion of psychological well-being, which reflects the realization of true self, positive relationships, human strengths, and virtues (Ryan and Deci, 2001; Ryff, 2014).

The conceptualization of existential issues is usually relevant to the framework of religion and spirituality (Park, 2005; Zinnbauer and Pargament, 2005). When people consider their global meanings about life and death, they often refer to the sacred aspect that is involved in both the definition of religiousness and spirituality (Pargament et al., 2005; Zinnbauer and Pargament, 2005).

The sacred includes concepts such as the divine, God, and the transcendent dimension, which provide an ultimate meaning to life and a sense of personal security and safety toward the unknown (Pargament et al., 2005).

The association between sacred and existential issues might be obvious for religious and spiritual people, but it could be less clear to those who do not attribute great importance to these topics in their lives (Pedersen et al., 2018). Thus, issues related to the global meaning of life should be considered in a broad secular way and not merely centered on a transcendent reality. Indeed, in light of a religious decline in Western societies (la Cour and Hvidt, 2010; Yu et al., 2017), beliefs in science or political ideology could play a role similar to that of religious beliefs for secular individuals (Farias et al., 2013).

Indeed, individual orientations toward a religious, spiritual, or secular perspective (or their possible overlap) do not take place in a social vacuum but rather depend on the cultural context in which a person lives (la Cour and Hvidt, 2010). For example, it has been shown that people living within a collectivist society tend to pursue a religious orientation in the existential experience by conforming to their own religious group, while people in the individualistic society tend to pursue a more secular orientation in the existential experience as a form of navigating personal uncertainty (Sullivan, 2013). In addition, the same individual could think about the global meanings in life in a religious, spiritual, and secular way, depending on his/her different phases of life (la Cour and Hvidt, 2010).

Several studies have attempted to develop measures concerning individual relationships with existential beliefs. For example, Thorne (1973) operationalized the person's existential status, which included concepts such as existential morale, existential vacuum, existence and destiny, and self-realization. Other scholars have assessed the degree to which people attribute meaning to and are aware of their own lives (Steger et al., 2006; Schulenberg et al., 2011; Richmond, 2015) or have measured individual factors related to existence, such as social and emotional loneliness, existential anxiety, death anxiety, and self-consciousness (Templer, 1970; Scheier and Carver, 1985; DiTommaso et al., 2004; Weems et al., 2004).

However, none of these measures directly assesses the degree to which people could be open to questioning themselves about existential issues, as the Existential Quest (EQ) scale (Van Pachterbeke et al., 2012) does. Perhaps the closest instruments are the Scale for Existential Thinking (Allan and Shearer, 2012) and the Religious Quest Scale (Batson and Schoenrade, 1991a,b). The former, like the EQ, investigates existential issues in a broad sense by assessing the frequency to which people think about these issues. The latter measures flexibility on existential issues but refers only to religious beliefs, and it was created to assess how people redefine their way of being religious as a consequence of contradictions and tragedies in life (Batson and Schoenrade, 1991a,b). Van Pachterbeke et al. (2012) developed the EQ to make a tool that assesses flexibility on existential issues available to all people, regardless of their being religious. To reach this goal, these authors introduced a new broad social-cognitive construct dealing with individual differences in their flexibility to change beliefs on core and universal issues, such as the ultimate meaning of life and the existence of transcendence. This form of open-mindedness could have positive implication at the societal level, because it is related to prosocial attitudes such as tolerance, altruism, and empathy. However, it can also have unfavorable implications at the individual level, because it could be related to feelings of uncertainty and anxiety, as for the religion quest attitude (Van Pachterbeke et al., 2012).

The EQ contains nine items assessing three different components, namely: a relative uncertainty regarding fundamental issues, a valuation of the doubt and questions surrounding these issues, and, eventually, openness to change (or the acknowledgment that one may change his or her own positions and attitudes across time).

In the original work, the authors assessed the factorial structure of the EQ scale and its discriminant validity by means of five studies involving several samples of students from Belgium and Germany and a sample of Belgian adults. As expected by Van Pachterbeke et al. (2012), EQ scores exhibited negative correlations with measures of closed-mindedness and positive correlations with measures related to prosocial attitudes and emotions. In particular, they found a negative correlation with the scores on the need for cognitive closure and Right-Wing Authoritarianism, and positive correlations with a measure of empathy and altruism. Religiousness was weakly correlated or uncorrelated with EQ scores across the studies, according to the hypothesis of independence of the EQ scores from religiousness. Furthermore, as they expected, a negative relationship of EQ scores with age was found (albeit weak). Lastly, as far as gender differences are concerned, no priory expectations were formulated and only in the sample of adults women scored higher than men. The dimensionality of the scale was evaluated by means of explorative factor analysis performed on one of the five studies and then replicated on the whole set of data from the five studies. In the single study, the authors found three factors that isolated religious items, doubt items, and the remaining items, respectively. Whereas on the pooled data a dominant factor of flexibility and a secondary factor dealing with flexibility in worldviews emerged. Supplementary analyses showing that the two factors provided the same pattern of associations with the majority of the variables included in the studies let the authors conclude that the scale could be conceived as unidimensional and that flexibility in worldviews and valuing doubt were facets of the same construct. The internal consistency of the nine items was acceptable ($\alpha = 0.74$).

The EQ has been applied in different fields of research. For example, Deak and Saroglou (2015, 2017) showed a positive correlation between EQ scores and measures of high tolerance toward moral questions, such as abortion, child euthanasia, gay adoption, and suicide. Furthermore, a negative correlation has been found with a measure of religious fundamentalism (Tapia Valladares et al., 2013), and a positive correlation has been shown with a measure of psychological well-being (Joshanloo, 2017). Finally, Sullivan (2013) showed that people belonging to an individualistic culture obtain higher scores on the EQ than those belonging to a collectivistic culture.

Given the relevance of the issues related to the EQ and, at the same time, the scarcity of instruments that investigate this quest, we consider it useful to deepen the psychometric characteristics of the EQ scale with an Italian sample.

AIMS

The aims of the study were threefold: (1) to examine the factor structure of the Italian adaptation of the EQ; (2) to test the measurement invariance separately across gender and age group; and (3) to assess the discriminant validity of the EQ scores with respect to measures of Right-Wing Authoritarianism (RWA) and the need for cognitive closure. Following Joshanloo (2017), we also tested the relation between EQ scores and a measure of psychological well-being. Furthermore, the relationship with gender, age, and religiousness was considered. To the best of our knowledge, this is the first attempt to confirm the psychometric properties of the EQ scale.

MATERIALS AND METHODS

Participants and Procedure

The participants were 291 Italian adults (64.3% female) aged 19 and 82 years (*M* = 37.0; *SD* = 14.6). Data collection occurred between April 2018 to June 2018; participants were recruited in the Northern part of Italy via a convenience sampling method through the dissemination of the questionnaire among university students attending degree courses in the field of social science (each student delivered some questionnaires to parents and/or acquaintances) (**Table 1**). The Ethic Committee of the University of Turin approved the study protocol. Participants took part voluntarily after giving their verbal consent to participate in the study. Respondents had to be at least 18 years of age to fill out the questionnaire.

Data were collected by means of a self-report pencil-and-paper questionnaire that took approximately 20 min to complete. A total of 98.9% of the respondents completed the questionnaire.

TABLE 1 | Characteristics as a percentage of the sample.

Characteristic	*n* = 291
Gender	
Female	64.3
Male	35.7
Employment status	
Students	27.1
Employed	62.2
Unemployed/retired	10.7
Education level completed	
Elementary school	1.0
Junior high school	6.2
High school	41.9
Bachelor's degree or higher	50.9

Measures

Existential Quest Scale (Van Pachterbeke et al., 2012)

The EQ was translated from English into Italian collegially by the authors and then was back translated by a native speaker. Participants were required to respond on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree). In the current study, Cronbach's alpha was 0.68. The original English items and the Italian adaptation of the EQ are reported in **Appendix 1**.

Right-Wing Authoritarianism Scale (Funke, 2005; Roccato et al., 2009)

The RWA is a 12-item self-report scale that assesses an overall authoritarianism attitude, rated on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). Cronbach's alpha found in the current study was 0.75.

Need for Cognitive Closure Scale-Brief Form (Pierro et al., 1995; Roets and Van Hiel, 2011)

We used a brief form (15 items) of the original scale of Webster and Kruglanski (1994), which assesses overall individual differences in cognitive closure. Participants responded on a 6-point scale ranging from 1 (not at all characteristic of me) to 6 (entirely characteristic of me). Cronbach's alpha found in the current study was 0.84.

The Mental Health Continuum-Short Form (Keyes, 2002; Petrillo et al., 2015)

The Mental Health Continuum-Short Form (MHC-SF) assesses three major dimensions of well-being: *psychological*, *social*, and *emotional*. Participants were asked to indicate how much of the time during the last month they functioned in a specific manner. Items were rated on a 6-point scale ranging from 0 (never) to 5 (always). The internal consistency found in the current study was good, Cronbach's alphas ranged from 0.77 to 0.82.

Religiousness

By means of principal component analysis, we calculated an index through three items created for the purpose of this study: "How much important is religion for you?," "Apart from weddings and funerals, how often do you attend mass or, if not Catholic, other religious rituals?," "How often do you attend the activities/initiatives of your religious group?." The items were rated on a 6-point scale ranging from 0 (not at all/never) to 5 (very much/more than once a week). In the current study, Cronbach's alpha was 0.84.

A brief list of sociodemographic items, including respondents' gender, age, and education, was also included.

We developed two versions of the questionnaire, presenting the EQ before and after the RWA and Need for Cognitive Closure Scale-Brief form (NFCS-BF) to prevent potential order effects. The MHC-SF was the first scale in both questionnaires.

Statistical Analyses

Imputation was performed using the expectation maximization (EM) method after it was verified that the missing values of the scales, ranging from 1.3 to 3.1%, were missing completely at random (MCAR) (Little, 1998).

We performed confirmatory factor analyses using MPLUS 7.3 (Muthén and Muthén, 1998-2015) to assess the factorial structure of the scale. According to the original study, we estimated a unidimensional model.

Because the data violated the multinormality condition [Mardia's multivariate omnibus test of skewness and kurtosis (2.26) = 125.60, $p < 0.001$], we used the Asparouhov and Muthén (2010) mean- and variance-adjusted ML (MLMV). As found by Maydeu-Olivares (2017), this estimation method has good properties in terms of the accuracy of standard errors and type I error in the presence of non-normal data. The following criteria were used to evaluate the acceptability of the goodness of fit of the model: root mean square error of approximation (RMSEA) $\leq$ 0.08; comparative fit index (CFI) $\geq$ 0.90; standardized root mean square residual (SRMR) $\leq$ 0.08 (Browne and Cudeck, 1993; Hu and Bentler, 1999). To assess measurement invariance, a multiple-group CFA (with gender and age as the grouping variables) was performed, and four increasingly restrictive models were estimated (Vandenberg and Lance, 2000). In the first model, all parameters were freely estimated across groups (configural invariance); in the second model, the loadings were assumed to be equal across groups (metric invariance); in the third model, both loadings and intercepts were constrained to be equal across groups (scalar invariance); and finally, in the fourth model, the residual variances were assumed to be equal across groups. The goodness of fit of each model was compared to that of the previous model (e.g., 2° vs. 1°; 3° vs. 2°). According to Chen (2007), the following changes in goodness-of-fit indices were considered indicative of a lack of invariance: ΔCFI ≤ -0.005; ΔRMSEA ≥ 0.010; regarding the SRMR, the cut-off was 0.025 for loading invariance and 0.005 for intercepts and uniqueness invariance.

The discriminant validity of the scale scores was tested by means of correlations (Pearson's *r*). Scale reliability was evaluated by means of the traditional Cronbach's α and by the Omega coefficient (ω, McDonald, 1978). As it is well known, α furnishes an unbiased estimate of reliability only when items conform to the essential tau-equivalence model under the Classical Test Theory (i.e., when items scores fit a unidimensional model in which the loadings are set to be equal and errors are uncorrelated). An appropriate alternative to α is the Omega coefficient (McDonald, 1999) that is based on the unidimensional model estimates and it is defined as the ratio between the variance due to the common factor and the variance of the total scale scores. In particular, the coefficient for measures with correlated errors was computed (Raykov and Marcoulides, 2016, p. 304).

All the analysis, except for CFAs, were performed with SPSS 25.0 (IBM SPSS Statistics, IBM Corporation).

RESULTS

Confirmatory Factor Analysis

The estimation of the one-factor model produced an unsatisfactory fit to the data: $\chi^2(27) = 150.1$, $p < 0.01$; RMSEA = 0.125 (90% CI = 0.11, 0.14); CFI = 0.639; and SRMR = 0.085.

To improve the model fit, we considered the contents of the items, looking for pairs of items that eventually shared part of their specificity. This examination identified three pairs of items that were more similar to each other than to the other elements of the scale. In detail, the pairings of items were as follows: items 1 and 7, the only items addressing the goal of life; items 2 and 9, the sole items related to the religious and spiritual sphere; and items 3 and 4, the unique items concerning the valorization of doubt. On the grounds of this consideration, with the support of the modification indices (MIs), the model was retested after the residuals of each item pair were correlated (1–7; 2–9; 3–4). The result of this model was satisfactory in terms of global fit indices: $\chi^2(24) = 51.5$, $p < 0.01$; RMSEA = 0.063 (90% CI = 0.04, 0.09); CFI = 0.919; and SRMR = 0.046.

As shown in **Table 2**, factor loadings (standardized values) were acceptable (>0.30), except for items 9 and 7, and all estimates were statistically significant ($p < 0.05$). The correlations between residuals were also not negligible (>0.30).

Measurement Invariance

The unidimensional model with three residual covariances obtained in the previous analysis was estimated in the multiple-group CFA to evaluate the degree of measurement invariance of EQ items across gender and age group.

The model imposing configural invariance across gender showed satisfactory fit values: $\chi^2(48) = 67.4$, $p < 0.05$; RMSEA = 0.053 (90% CI = 0.01, 0.08); CFI = 0.939; and SRMR = 0.055. However, a close examination of the loadings showed that, in the group of men, the loading of item 7 was not statistically significant (0.05; $p = 0.84$). Thus, we excluded item 7 from the analysis of gender invariance, and this exclusion reduced the number of residual covariances to be estimated: the covariance between items 1 and 7 was no longer a model parameter. As shown in **Table 3**, on the remaining eight items, all the models – from the one that imposes equality of the loading pattern (configural) to the one that imposes equality of all item parameters (uniqueness

TABLE 2 | Standardized loadings for one-factor confirmatory model of Existential Quest Scale (*n* = 291).

Item	Loading
1. Today, I still wonder about the meaning and goal of my life	0.43
2. My attitude toward religion/spirituality is likely to change according to my life experiences	0.39
3. Being able to doubt about one's convictions and to reappraise them is a good quality	0.45
4. In my opinion, doubt is important in existential questions	0.50
5. My way of seeing the world is certainly going to change again	0.71
6. My opinion varies on a lot of subjects	0.52
7. I know perfectly well what the goal of my life is*	0.16
8. Years go by, but my way of seeing the world doesn't change*	0.41
9. I often reappraise my opinion on religious/spiritual beliefs	0.24

**Item reverse-coded. Model estimates include three correlations between residuals: 0.48 (items 2 and 9); 0.38 (items 3 and 4); and 0.30 (items 1 and 7). All estimates are statistically significant at p < 0.05.*

TABLE 3 | Measurement invariance of the EQ scale.

Models across gender	χ^2	*df*	RMSEA	CFI	SRMR	$\Delta\chi^2$	Δ*df*	ΔRMSEA	ΔCFI	ΔSRMR
Males	25.3	18	0.062	0.928	0.060	–	–	–	–	–
Females	20.3	18	0.026	0.988	0.036	–	–	–	–	–
1. Configural	45.6	36	0.043	0.967	0.046	–	–	–	–	–
2. Metric$_a$	53.9	45	0.037	0.969	0.051	8.73	9	−0.006	0.002	0.005
3. Scalar$_a$	60.0	52	0.032	0.972	0.055	5.76	7	−0.005	0.003	0.004
4. Uniquenesses$_a$	68.8	60	0.032	0.970	0.059	9.50	8	0.000	−0.002	0.004
Models across age										
Adults (aged ≥ 31 years)	20.4	18	0.030	0.984	0.043	–	–	–	–	–
Young adults (aged < 31 years)	28.8	18	0.065	0.923	0.049	–	–	–	–	–
1. Configural	48.5	36	0.049	0.957	0.046	–	–	–	–	–
2. Metric$_a$	59.5	45	0.047	0.951	0.058	11.74	9	−0.002	−0.006	0.012
3. Scalar$_a$	75.4	52	0.056	0.920	0.070	18.11*	7	0.009	−0.031	0.012
3a. Scalar$_{a,b}$	65.3	50	0.046	0.948	0.062	5.92	5	−0.001	−0.003	0.004
4. Uniquenesses$_a$	79.8	58	0.051	0.926	0.077	15.64*	8	0.005	−0.022	0.015
4a. Uniquenesses$_{a,c}$	70.5	56	0.042	0.951	0.066	6.11	6	−0.004	0.003	0.004

*RMSEA, root mean square error of approximation; CFI, comparative fit index; SRMR, standardized root mean square residual. [a]The error covariance between items 2 and 9 and between items 3 and 4 was constrained to be equal across groups; [b]Free intercept on items 8 and 1; [c]Free uniqueness on items 8 and 1. *p < 0.05.*

invariance) – showed excellent fit to the data. The non-significant difference in χ^2 ($\Delta\chi^2$) and the very small change in RMSEA, CFI, and SRMR obtained in each of the comparisons lend support to the idea that the EQ items exhibit full measurement invariance across gender.

With the aim of assessing measurement invariance with respect to age, two groups were formed using the median of the sample (31 years) as a cut-off (young adults, N = 142; adults, N = 149). The fit of the configural model on the nine items of the scale was adequate [$\chi^2(48) = 66.8$, $p < 0.05$; RMSEA = 0.052 (90% CI = 0.01, 0.08); CFI = 0.939; SRMR = 0.053]. However, as in the gender group analyses, the loading of item 7 was not statistically significant; in this case, it was not statistically significant in either of the two groups (young adults: 0.28, $p = 0.14$; adults: 0.07, $p = 0.74$). Thus, we also dropped item 7 in this analysis. As shown in **Table 3**, the configural and metric models provided excellent fit to the data. In terms of changes in the fit measures, in the metric invariance model, only ΔCFI was slightly above the cut-off, but we did not consider this lack of fit to be problematic because all the other changes in fit indices were small. The imposition of the equality of the intercepts resulted in a remarkable change in both the CFI and SRMR. To evaluate whether partial scalar invariance was tenable, we examined the MIs relative to the item intercepts, and we relaxed the equality constraint on the item intercept associated with the largest MI, one at a time, until the changes in the fit indices with respect to the metric invariance model were negligible. After the intercept equality constraint on items 8 and 1 was removed, changes in the fit indices were very small. Regarding the uniqueness invariance, both ΔCFI and ΔSRMR were outside the range. The inspection of MI suggested the removal of the equality constraint from the uniqueness of items 8 and 1, thus leading to a satisfactory model fit. These two items were not invariant across age groups, and both items exhibited lower intercept and greater uniqueness in the adult sample than in the younger sample.

Discriminant Validity

To correlate EQ scores with those of the other scales, a total mean score of flexibility was computed. In light of the results obtained above, item 7 was excluded from the computation (means and standard deviations of EQ items are shown in **Appendix 2**).

As reported in **Table 4**, EQ scores showed a moderate negative correlation with RWA scores and a weak negative correlation with NFCS-BF scores. Flexibility scores were not correlated with well-being scores, neither with subscales nor with total scores.

Regarding the religiousness index, no correlation was found, and no relationship emerged with respect to gender. Flexibility scores were negatively correlated with age, although the correlation was weak.

Internal Consistency

For the 8-items scale, Cronbach's α was 0.70 and McDonald's ω was 0.61, meaning that 61% of the total score variance was due to the common latent factor. The difference between α and ω was mainly due to the presence of correlated errors. In fact, when omega was computed including the error covariances among the systematic part at the numerator of the formula:

$$\frac{(\sum \lambda_i)^2 + 2^* \sum \sigma_{i,j}}{(\sum \lambda_i)^2 + 2^* \sum \sigma_{i,j} + \sum \sigma_i^2},$$

the value (0.69) was very close to that of α.

DISCUSSION

The study investigated the psychometric properties of the EQ across an Italian sample. The results supported the unidimensionality of the scale, in line with the findings of the original study of Van Pachterbeke et al. (2012). More specifically, scale scores were essentially unidimensional, because the presence of some error covariances signals that there are some

TABLE 4 | Summary of intercorrelations for scores on the EQ and the other study variables.

	1	2	3	4	5	6	7	8	9	10
1. EQ	–									
2. RWA	−0.38**	–								
3. NFCS	−0.14*	0.39**	–							
4. MHC	−0.06	0.07	−0.15**	–						
5. EWB	−0.04	−0.01	−0.17**	0.79**	–					
6. SWB	−0.01	0.05	−0.16**	0.86**	0.53**	–				
7. PWB	−0.09	0.10	−0.09	0.91**	0.65**	0.63**	–			
8. Religiousness	−0.08	0.36**	−0.02	0.27**	0.15**	0.32**	0.21**	–		
9. Gender (0 = M)	0.07	−0.04	0.01	−0.01	0.03	−0.03	0.02	0.02	–	
10. Age	−0.21**	0.24**	0.24**	−0.01	0.05	−0.11	0.09	0.01	−0.07	–

EQ, Existential Quest Scale without item 7; RWA, Right-wing Authoritarianism; NFCS, Need for Cognitive Closure Scale-Brief form; MHC, Mental Health Continuum-Short form (total score); PWB, Psychological well-being; SWB, social well-being; EWB, emotional well-being. *$p < 0.05$; **$p < 0.01$.

secondary dimensions. However, this result is consistent with the intention of the proposers of the scale to develop a broad measure of flexibility by using a set of items "that do not merely paraphrase each other, including items that address the different components of the quest orientation" (Van Pachterbeke et al., 2012, p. 3). The presence of more than one item for each component created undesired covariation between items (as for the two items about religious beliefs and the two relative to evaluating doubt). At the same time, the number of items per component was too small to substantiate the presence of a general factor and some content-related factors (group factors).

One item ("I know perfectly well what the goal of my life is") performed poorly both in the factor analysis conducted on the whole sample and in the measurement invariance tests across gender and age groups. This result was in line with those of previous studies that found this item to be a poor indicator of existential flexibility (Van Pachterbeke et al., 2012; Joshanloo, 2017). In light of these considerations, we do not advise the consideration of item 7 in the EQ.

The 8-item scale revealed full measurement invariance across gender, reflecting that there are no differences in the Italian sample between males and females in the EQ factorial structure, while partial invariance emerged across age groups because two items (items 1 and 8) differed both in terms of intercept and residual variance across younger and older adults. Considering the contents and formulations of these items, some considerations can be formulated. It is plausible that being uncertain about the meaning of life (item 1) has different implications for younger and older adults. For younger people more than for older people, it could be a positive aspect associated with the openness to new experiences, whereas for older than for younger people, it could have a negative, depressive connotation. Similarly, the significance of the item about changing the way of seeing the world (item 8) may have a different meaning according to the age of respondents, especially because this item refers to change occurring "over the years".

In line with the original study (Van Pachterbeke et al., 2012), EQ scores showed good discriminant validity in terms of their correlation with RWA and NFCS-BF scores. High flexibility in EQ was associated with the tendency to be autonomous with respect to norms (low RWA) and to be less cognitively rigid (low need for cognitive closure). Furthermore, consistent with the literature, we found that younger people are more flexible with respect to existential questions than older people are.

As concern internal consistency of the total scale score, α-value was similar to that obtained in the original study (α = 0.74), and quite higher than the value of omega. Thus, our results are coherent with those of Gu et al. (2013) who found that α tends to treat correlated error variance as true variance and thus inflates the estimate of reliability. The value of omega was low, but this result does not imply necessarily that the scale is heavily affected by random error variation. The low value could be mainly due to item specificity that is, the influence of factors that are specific for each item. The item specificity is a source of systematic variation that could be considered a component of the "true" variance, depending on the definition of reliability the researcher is adopting. Even if in a single administration, as in the present study, it is not possible to distinguish between random variation and items specificity, we can conjecture that specificity is a not negligible component of EQ scores because, as stated above, EQ was intended as a broad measure of flexibility.

In summary, the present study assessed for the first time the factorial structure of the EQ by means of a confirmatory approach. The study provided some evidence of measurement invariance across gender and age and showed that the Italian version of the scale presents satisfactory psychometric properties. Nonetheless, this study is not exempt from some limitations. Firstly, because of the type of sampling method employed, the participants were not representative of the Italian population, with an over-representation of women and high educated people. Secondly, although the number of participants was adequate to perform the intended analyses, it did not allow for the formation of more than two age groups, thus limiting the exploration of the functioning of the items according to age. Moreover, it did not allow splitting the sample and performing both exploratory and confirmatory factor analyses. The exploratory approach with a bi-factor rotation could be useful in further exploration of the factorial structure of the scale, because it allows modeling a general factor and two or more group factors related to the content components of the scale. We could not estimate a confirmatory bi-factor model because a minimum of three indicators for each group factor is request. Moreover, further researches aimed at assessing EQ reliability by means of a test–retest design are recommendable in order to assess how much EQ total score is affected by random variation (McCrae, 2015). Finally, although promising, we have collected data in a predominantly Catholic country, so it is necessary to investigate the properties of this scale in other countries with different cultural and religious traditions.

CONCLUSION

The new construct and the relative scale developed by Van Pachterbeke et al. (2012) could be used in several field of psychology (social, clinical, developmental) as it deals with issues that more or less involve all human beings in every period of life since the development of abstract and critical thinking.

The EQ scale may represent a useful tool to better understand how people experience different perspectives in Western societies characterized by the coexistence of different cultures and religions. Assessing individual differences in their flexibility on existential issues could help to understand why some people are willing to accept the presence of people with different cultures and/or religions and others tend to do not tolerate the contradiction due to the multicultural presence.

At the individual level, being more or less an existential quester could be related to personal well-being. In contrast to previous studies that reported a positive correlation between the two measures (Joshanloo, 2017), our results failed to find a significant relationship between the EQ and individuals' well-being. Indeed, high flexibility with respect to the EQ could combine with emotional instability and anxiety, as claimed in the original study (Van Pachterbeke et al., 2012). In other words, future studies could aim to disambiguate the positive or negative contribution of such flexibility in individuals' lives, as flexibility may help manage stressful situations such as disabling illness (la Cour, 2008) but could also be related to existential anxiety and an increase in risky behaviors during adolescence (Carter et al., 2013).

DATA AVAILABILITY

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethic Committee of the University of Turin (code 10039). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MR, ST, SG, and AM conceived the study. MR and ST did the analyses. MR wrote the manuscript. All authors discussed the results together and contributed to the final manuscript, doing critical revisions and giving suggestions, and approved the submitted version of the manuscript.

FUNDING

ST, SG, and AM were supported by the University of Turin (Ricerca Scientifica Finanziata dall'Università).

REFERENCES

Allan, B. A., and Shearer, C. B. (2012). The Scale for existential thinking. *Int. J. Transpers. Stud.* 31, 21–37. doi: 10.24972/ijts.2012.31.1.21

Asparouhov, T., and Muthén, B. (2010). *Simple Second Order Chi-Square Correction Scaled Chi-Square Statistics (Technical Appendix)*. Los Angeles,CA: Muthén & Muthén.

Batson, C., and Schoenrade, P. (1991a). Measuring religion as quest: 1) validity concerns. *J. Sci. Study Relig.* 30, 416–429. doi: 10.2307/1387277

Batson, C., and Schoenrade, P. (1991b). Measuring religion as quest: 2) reliability concerns. *J. Sci. Study Relig.* 30, 430–447. doi: 10.2307/1387278

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen, and J. S. Long, (London: Sage), 132–162.

Carter, J., Berman, S. L., Marsee, M. A., and Weems, C. F. (2013). Identity exploration, commitment, and existential anxiety as predictors of the forms and functions of aggression. *Identity* 13, 348–367. doi: 10.1080/15283488.2013.780975

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

Deak, C., and Saroglou, V. (2015). Opposing abortion, gay adoption, euthanasia, and suicide. *Arch. Psychol. Relig.* 37, 267–294. doi: 10.1163/15736121-12341309

Deak, C., and Saroglou, V. (2017). Terminating a child's life? Religious, moral, cognitive, and emotional factors underlying non-acceptance of child euthanasia. *Psychol. Belg.* 57, 59–76. doi: 10.5334/ pb.341

DiTommaso, E., Brannen, C., and Best, L. A. (2004). Measurement and validity characteristics of the short version of the social and emotional loneliness scale for adults. *Educ. Psychol. Meas.* 64, 99–119. doi: 10.1177/0013164403258450

Farias, M., Newheiser, A., Kahane, G., and de Toledo, Z. (2013). Scientific faith: belief in science increases in the face of stress and existential anxiety. *J. Exp. Soc. Psychol.* 49, 1210–1213. doi: 10.1016/j.jesp.2013.05.008

Funke, F. (2005). The dimensionality of right-wing authoritarianism: lessons from the dilemma between theory and measurement. *Polit. Psychol.* 26, 195–218. doi: 10.1111/j.1467-9221.2005.00415.x

Gu, F., Little, T. D., and Kingston, N. M. (2013). Misestimation of reliability using coefficient alpha and structural equation modeling when assumptions of tau -equivalence and uncorrelated errors are violated. *Methodology* 9, 30–40. doi: 10.1027/1614-2241/a000052

Hu, L., and Bentler, P. M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Modeling* 26, 1–55. doi: 10.1080/10705519909540118

Joshanloo, M. (2017). Factorial/Discriminant validity and longitudinal measurement invariance of MHC-SF in Korean young adults. *Curr. Psychol.* 1–7. doi: 10.1007/s12144-017-9742-1

Keyes, C. L. M. (2002). The mental health continuum: from languishing to flourishing in life. *J. Health Soc. Behav.* 43, 207–222. doi: 10.2307/3090197

la Cour, P. (2008). Existential and religious issues when admitted to hospital in a secular society: patterns of change. *Ment. Health Relig. Cult.* 11, 769–782. doi: 10.1080/13674670802024107

la Cour, P., and Hvidt, N. C. (2010). Research on meaning-making and health in secular society: secular, spiritual and religious existential orientations. *Soc. Sci. Med.* 71, 1292–1299. doi: 10.1016/j.socscimed.2010.06.024

Little, R. J. A. (1998). A test of missing completely at random for multivariate data with missing values. *J. Am. Stat. Assoc.* 83, 1198–1202. doi: 10.1080/01621459.1988.10478722

Maydeu-Olivares, R. (2017). Maximum likelihood estimation of structural equation models for continuous data: standard errors and goodness of fit. *Struct. Equ. Modeling* 24, 383–394. doi: 10.1080/10705511.2016.1269606

McCrae, R. R. (2015). A more nuanced view of reliability: specificity in the trait hierarchy. *Pers. Soc. Psychol. Rev.* 19, 97–112. doi: 10.1177/1088868314541857

McDonald, R. P. (1978). Generalizability in factorable domains: "domain validity and generalizability". *Educ. Psychol. Meas.* 38, 75–79. doi: 10.1177/001316447803800111

McDonald, R. P. (1999). *Test Theory: A Unified Approach*. Mahwah, NJ: Erlbaum.

Muthén L. K. and Muthén (1998-2015). *Mplus User's Guide. Seventh Edition*. Los Angeles, CA: Muthén & Muthén. doi: 10.1177/001316447803800111

Pargament, K. I., Magyar-Russell, G. M., and Murray-Swank, N. A. (2005). The sacred and the search for significance: religion as a unique process. *J. Soc. Issues* 61, 665–687. doi: 10.1111/j.1540-4560.2005.00426.x doi: 10.1111/j.1540-4560.2005.00426.x

Park, C. L. (2005). "Religion and meaning," in *Handbook of the Psychology of Religion and Spirituality*, eds R. F. Paloutzian, and C. L. Park, (New York, NY: Guilford), 295–314.

Pedersen, H. F., Birkeland, M. H., Jensen, J. S., Schnell, T., Hvidt, N. C., Sørensen, T., et al. (2018). What brings meaning to life in a highly secular society? A study on sources of meaning among Danes. *Scand. J. Psychol.* 59, 678–690. doi: 10.1111/sjop.12495

Petrillo, G., Capone, V., Caso, D., and Keyes, C. L. M. (2015). The mental health continuum–short form (MHC–SF) as a measure of well-being in the Italian context. *Soc. Indic. Res.* 121, 291–312. doi: 10.1007/s11205-014-0629-3

Pierro, A., Mannetti, L., Converso, D., Garsia, V., Miglietta, A., Ravenna, M., et al. (1995). Caratteristiche strutturali della versione italiana della scala di bisogno di chiusura cognitiva (di Webster and Kruglanski). *TPM* 2, 125–141.

Raykov, T., and Marcoulides, G. A. (2016). Scale reliability evaluation under multiple assumption violations. *Struct. Equ. Modeling* 23, 302–313. doi: 10.1080/10705511.2014.938597

Richmond, M. M. (2015). *Development of an Instrument Measuring Existential Authenticity*. Ph.D. thesis, University of Cincinnati, Cincinnati, OH.

Roccato, M., Mirisola, A., and Chirumbolo, A. (2009). La rilevazione empirica dell'autoritarismo di destra: un contributo all'adattamento italiano della scala Funke. *Psicol. Soc.* 1, 157–174.

Roets, A., and Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Pers. Individ. Dif.* 50, 90–94. doi: 10.1016/j.paid.2010.09.004

Ryan, R. M., and Deci, E. L. (2001). On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Ann. Rev. Psychol.* 52, 141–166. doi: 10.1146/annurev.psych.52.1.141

Ryff, C. D. (2014). Psychological well-being revisited: advances in the science and practice of eudaimonia. *Psychother. Psychosom.* 83, 10–28. doi: 10.1159/000353263

Scheier, M. F., and Carver, C. S. (1985). The self-consciousness scale: a revised version for use with general populations. *J. Appl. Soc. Psychol.* 15, 687–699. doi: 10.1111/j.1559-1816.1985.tb02268.x

Schulenberg, S. E., Schnetzer, L. W., and Buchanan, E. M. (2011). The purpose in life test-short form: development and psychometric support. *J. Happiness Stud.* 12, 861–876. doi: 10.1007/s10902-010-9231-9

Steger, M. F., Frazier, P., Oishi, S., and Kaler, M. (2006). The meaning of life questionnaire: assessing the presence of and search for meaning in life. *J. Couns. Psychol.* 53, 80–93. doi: 10.1080/00223891.2013.765882

Sullivan, D. (2013). *Disorientation-Avoidant and Despair-Avoidant Cultures*. Ph.D. thesis, University of Kansas, Lawrence, KS.

Tapia Valladares, J., Rojas Carvajal, M., and Villalobos García, M. (2013). Religious fundamentalism among Costa Rican University: students political conservatism and spirituality without religion. *Rev. Cienc. Soc.* 1, 115–135. doi: 10.15517/RCS.V0I139.11357

Templer, D. I. (1970). The construction and validation of a death anxiety scale. *J. Gen. Psychol.* 82, 165–177. doi: 10.1080/00221309.1970.9920634

Thorne, F. C. (1973). The existential study: a measure of existential status. *J. Clin. Psychol.* 29, 387–392. doi: 10.1002/1097-4679(197310)29

Van Pachterbeke, M., Keller, J., and Saroglou, V. (2012). Flexibility in existential beliefs and worldviews: introducing and measuring existential quest. *J. Individ. Dif.* 33, 2–16. doi: 10.1027/1614-0001/a000056

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Webster, D. M., and Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *J. Pers. Soc. Psychol.* 67, 1049–1062.

Weems, C. F., Costa, N. M., Dehon, C., and Berman, S. L. (2004). Paul Tillich's theory of existential anxiety: a preliminary conceptual and empirical examination. *Anxiety Stress Coping* 17, 383–399. doi: 10.1080/10615800412331318616

Yu, C. H., Reimer, D., Lee, A., Snijder, J., and Lee, H. S. (2017). A triangulated and exploratory study of the relationships between secularization, religiosity, and social wellbeing. *Soc. Indic. Res.* 131, 1103–1119. doi: 10.1007/s11205-016-1290-9

Zinnbauer, B. J., and Pargament, K. I. (2005). "Religion and sprituality," in *Handbook of the Psychology of Religion and Spirituality*, eds R. F. Paloutzian, and C. L. Park, (New York, NY: Guilford), 21–42.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX 1

Existential Quest (English Version in Brackets)

1. Ad oggi, mi pongo ancora delle domande sul significato e lo scopo della mia vita [Today, I still wonder about the meaning and goal of my life].
2. Sulla base delle esperienze della mia vita, il mio approccio verso la religione/spiritualità probabilmente cambierà [My attitude toward religion/spirituality is likely to change according to my life experiences].
3. Mettere in dubbio le proprie convinzioni e rivalutarle è una caratteristica positiva [Being able to doubt about one's convictions and to reappraise them is a good quality].
4. Penso che il dubbio abbia un ruolo importante nelle domande esistenziali [In my opinion, doubt is important in existential questions].
5. Il mio modo di vedere il mondo sicuramente cambierà ancora [My way of seeing the world is certainly going to change again].
6. La mia opinione su molti argomenti varia [My opinion varies on a lot of subjects].
7. Ho ben presente qual è lo scopo della mia vita [I know perfectly well what the goal of my life is](*).
8. Passano gli anni ma il mio modo di vedere il mondo non-cambia [Years go by, but my way of seeing the world doesn't change](*).
9. Spesso rivaluto la mia opinione sulle credenze religiose/spirituali [I often reappraise my opinion on religious/spiritual beliefs]

(*) reverse-scored item.

APPENDIX 2

Descriptives of the Existential Quest Scale

Item	*M*	*SD*	Skew	Kurt
1	5.21	1.81	−0.90	−0.22
2	3.43	1.93	0.26	−1.09
3	5.95	1.31	−1.20	0.71
4	5.88	1.26	−1.23	1.31
5	5.55	1.34	−1.04	1.11
6	4.78	1.53	−0.24	−0.84
7*	3.85	1.87	0.07	−1.07
8*	5.05	1.71	−0.82	−0.21
9	3.00	1.81	0.61	−0.81
Scale scores	4.74	0.87	−0.23	−1.18
Scale scores[a]	4.85	0.91	−0.31	−0.07

**Reverse-scored item. [a]Scale scores without item 7.*

Confirmatory Factor Analysis of the Enriched Life Scale Among US Military Veterans

Caroline M. Angel[1,2,3]*, Mahlet A. Woldetsadik[4]*, Justin T. McDaniel[5], Nicholas J. Armstrong[3], Brandon B. Young[1,2,6], Rachel K. Linsner[3] and John M. Pinter[1]

[1] Team Red, White & Blue, Alexandria, VA, United States, [2] Reintegrative Health Initiative, Westfield, NJ, United States, [3] Institute for Veterans and Military Families, Syracuse University, Syracuse, NY, United States, [4] Pardee RAND Graduate School, Santa Monica, CA, United States, [5] Department of Public Health and Recreation Professions, Southern Illinois University, Carbondale, IL, United States, [6] Tennyson Center for Children, Denver, CO, United States

OPEN ACCESS

Edited by:
Elisa Pedroli,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Sonja Heintz,
University of Zurich, Switzerland
Ali Montazeri,
Iranian Institute for Health Sciences
Research, Iran

***Correspondence:**
Caroline M. Angel
caroline.angel@teamrwb.org
Mahlet A. Woldetsadik
mahal.meare@gmail.com

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: 21 March 2019
Accepted: 10 September 2019
Published: 11 October 2019

Citation:
Angel CM, Woldetsadik MA,
McDaniel JT, Armstrong NJ,
Young BB, Linsner RK and Pinter JM
(2019) Confirmatory Factor Analysis
of the Enriched Life Scale Among US
Military Veterans.
Front. Psychol. 10:2181.
doi: 10.3389/fpsyg.2019.02181

The Enriched Life Scale (ELS) is a 40-item measure developed by the military veteran service organization, Team Red, White & Blue (RWB), to systematically capture and quantify the lived experiences of military veterans transitioning to civilian life. As Team RWB's mission is to "enrich veterans' lives," veterans who conceived of and co-developed the ELS as a psychometric instrument defined what an "enriched life" would entail. Exploratory factor analysis (EFA) of the ELS revealed a five-factor structure capturing the domains of: physical health, mental health, genuine relationships, sense of purpose, and engaged citizenship. The goal of the current study was to use confirmatory factor analysis to validate the factor structure of the ELS in a sample of veterans not affiliated with Team RWB. We also sought to explore convergent validity with the Military to Civilian Questionnaire, a measure of military to civilian reintegration challenges. Five hundred and twenty-nine veterans participated in the study. We estimated three models, one-factor, four-factor, and five-factor model via maximum likelihood estimation with robust Huber-White standard errors. The five-factor model showed the best fit to the data (RMSEA = 0.05, CFI = 0.90, TLI = 0.90, SRMR = 0.06). Additionally, the five-factor model demonstrated convergent and discriminant validity, as well as internal consistency reliability (genuine relationships, $\alpha = 0.90$; sense of purpose, $\alpha = 0.93$; engaged citizenship, $\alpha = 0.89$; mental health, $\alpha = 0.88$; and physical health, $\alpha = 0.78$). Overall, the ELS is a valid and reliable measure of veteran enrichment and could potentially be used in conjunction with diagnostic instruments that capture strain-related transition challenges (to include mental health disorders) to capture post-military service wellbeing.

Keywords: Enriched Life Scale, confirmatory factor analysis, veteran, wellbeing, Team Red, White & Blue, psychometric assessment

INTRODUCTION

Military veterans must navigate a range of challenges in their transition to civilian life. While the transition from service member to veteran is primarily characterized by resilience, many veterans experience lasting physical, psychological, and social problems related to military service and reintegration (Angel, 2016; Elnitsky et al., 2017; Mobbs and Bonanno, 2018). Team Red, White

& Blue (RWB) was founded in 2010 to offset service-related reintegration stressors by providing opportunities for veterans to connect with service-connected peers and civilian community members. Over 200 national chapters create local and consistent opportunities for members to participate in physical, social, leadership, and volunteering activities. In 2018, Team RWB's membership reached over 153,000 members and over 2,000 volunteer leaders created 38,000 Team RWB events[1]. The mission of Team RWB is to "enrich veterans' lives" and the foundational veteran thought leaders spent years developing this theoretical model of engagement and defining what it means to "enrich" a life. Leaders ultimately defined an "enriched life" as having physical, mental, and emotional health; genuine relationships comprised of close, best-friend types of relationships within a broader social network; and a sense of purpose, which included an individual sense of purpose, shared purpose, and positive role identity (Angel et al., 2018a).

Team RWB was founded in 2010 by Army Captain, Michael S. Erwin, who was studying positive psychology principles under the field's co-founder, Christopher Peterson. Positive psychology focuses on "what goes right in life" (Seligman and Csikszentmihalyi, 2000); Team RWB was established to connect transitioning veterans to their community through activities that supported physical activity and helped them develop and maintain personal and community connections (Angel and Armstrong, 2016). With increasing negative health behaviors and weight gain as major issues affecting veterans along with the loss of camaraderie, sense of purpose, and shared mission with others, Team RWB was filling a gap by offering a new approach to supporting transitioning veterans to their communities (Angel et al., 2018a). While Team RWB was leveraging the principles of positive psychology for community dwelling veterans, the movement of positive psychology had just begun to rise in the Army itself. In 2008, the Army implemented the Comprehensive Soldier Fitness Program, designed to increase active duty soldiers' psychosocial and positive performance through assessment and training. As physical fitness tests were already routinely in place, Army leaders were proactively developing soldier psychosocial resilience thereby hoping to decrease psychological disorders as a result of military service (Cornum et al., 2011). As *resilience* became the focus of the Army's positive psychology training program, Team RWB leaders purposely avoided language reminiscent of active duty service, which they believed would be off putting to new members who were recently transitioned out of the service, and may wish to avoid that reminder. "Enriching lives," ultimately most resonated with Team RWB's founder more so than other concepts of well-being, to which it is highly related (Angel and Armstrong, 2016).

In the extant literature, the concept of an "enriched life" is theoretically related to constructs such as well-being, life satisfaction, and flourishing (Angel et al., 2018a). We have previously described how conceptualizations of "veteran wellness" broadly defined as satisfactory function in the areas of personal relationships, health, fulfillment of material needs, and having a sense of purpose is applicable to veterans and civilians alike (Angel et al., 2018a). More traditional conceptualizations of well-being, however, have traditionally neglected the physical health component. Ryff's (2018) foundational definition of "well-being" was formulated based upon the philosophical tenets first articulated by Aristotle and developed by psychologists from clinical, developmental, humanistic, existential, and social perspectives. Ryan and Deci (2001) defined well-being as optimal psychological functioning and experience organized by two central perspectives: hedonic and eudaimonic well-being. The hedonic approach focuses on pleasure seeking and pain avoidance for body and mind while the eudaimonic approach focuses on meaning and self-actualization. In the eudaimonic tradition, Ryff developed a theory-guided measure of psychological well-being. The widely used measure, the Ryff Scales of Psychological Well-Being assessed six constructs: self-acceptance (positive attitude toward the self), positive relations with others (warm, satisfying, trusting relationships with others), autonomy (self-determining/independent), environmental mastery (competence in managing the environment), purpose in life (direction and meaning in life), and personal growth (feelings of continued development) (Ryff, 1989). Veteran conceptualization of their own well-being is aligned to eudaimonic approaches, integrating a sense of purpose and opportunities to serve others through volunteering and leading others as key components.

"Life satisfaction" has been deemed a cognitive component of subjective well-being, described as a general self-appraisal of one's own quality of life (Pavot and Diener, 2009). It's most widely used measure, the Satisfaction with Life Scale (Diener et al., 1985), is unidimensional, capturing the factor of "life satisfaction" which is theoretically related to an enriched life. Finally, the concept of "flourishing" has been defined as having positive emotion, engagement, relationships, meaning and accomplishment (Seligman, 2011). While more recent conceptualizations of flourishing published following the development of the Enriched Life Scale (ELS) have included references to positive physical health (VanderWeele, 2017), traditional conceptualizations of flourishing have primarily overlooked physical health as a key component.

Team RWB leaders explored a variety of existing instruments prior to the development of the ELS. Scales that have measured well-being have trended to capture between one to six constructs on the dimensions of well-being: global well-being, social well-being, physical well-being, spiritual well-being, activities and functioning, and personal circumstances, and run between five and one hundred or more items (Linton et al., 2018). Linton et al.'s (2018) review of 99 self-report measures for assessing wellbeing in adults describes these instruments in depth. While Team RWB leaders admittedly did not examine every instrument reviewed by Linton et al. (2018) prior to the development of the ELS in 2014, they believed the original enrichment equation (five constructs to include physical health; mental health; emotional health; genuine relationships; and sense of purpose) would need to measure all domains that they felt captured veterans' lived experience of an enriched life and was detailed enough to provide information back to the organization so that Team RWB leaders could actively engage members through specific,

[1] https://www.teamrwb.org/reports/annual-report-2018/

needs-driven (potentially individualized) activities. Therefore, driven by their operational experience of designing and deploying survey instruments in a non-profit membership environment, for which the ELS was originally developed, they hypothesized that the instrument should be between 25 and 45 items. Existing instruments considered, like the Ryff Scales of Psychological Well Being (Ryff, 1989), the Perma Profiler (Butler and Kern, 2016), the Flourishing Scale (Diener et al., 2010), were deemed too narrow in scope theoretically or too short to adequately capture what veteran leaders felt defined an "enriched life". Other instruments provided simple yes/no checklists yielding too limited information to provide operationally useful feedback (Linton et al., 2018). Additionally, at least two widely used scales, the Conner Davidson Resilience Scale (Green et al., 2014), and the Perma Profiler (Butler and Kern, 2016) have demonstrated new factor structures differing from the original when tested in veteran populations (Umucu et al., 2019).

Veteran leaders and consulting academics also considered the translational capabilities of existing measures. They viewed the translation of the constructs of other popular instruments to a broader lay-person public health communication strategy as limited. The concepts themselves are semantically representative of academic terminology and would be lost on an audience unfamiliar with such discipline-specific terms (for example, "environmental mastery"). Often they found the terminology lacking cultural congruity to veteran serving community based organizations, in which communications are more generally guided by marketing, development, and personal relations professionals than researchers or clinicians. Even the U.S. Army developed Global Assessment Tool, an assessment of soldier psychosocial fitness tailored to the Comprehensive Soldier Fitness Program, did not assess physical health component, which is fundamental to Team RWB's mission; given its 105 item length, it could not feasibly be administered to newly joining members of the community based veteran service organization.

Therefore, Team RWB veteran thought leaders and social scientists spent three years (2014–2017) developing the ELS (Team Red White Blue, 2017), which was finalized as a 40-item instrument in 2017 (Team Red White Blue, 2017; Angel et al., 2018b). The need for an instrument with valid and reliable psychometric purposes was driven by Team RWB's desire to be accountable and transparent to key stakeholders (members, funders, supporters) in their articulation and measurement of the impact of their programs in achieving their stated mission. Additionally, the ability to provide a veteran-developed assessment tool which placed veterans' needs and lived experiences of transition from military to civilian life as the guiding voices in determining successful transition filled an assessment and research gap. It also permitted the development of an instrument that could feasibly be administered to thousands of newly joining Team RWB members, which Team RWB is currently exploring.

Preliminary psychometric properties were established for the 40-item ELS in a sample of 1,187 military veterans and 598 civilians, all members of Team RWB (Angel et al., 2018b). The theoretical model of an "enriched life" was mostly validated, with the exception that the hypothesized construct, "emotional health" did not emerge as a stand-alone construct. Instead, items originally written to reflect the definition and measurement of "emotional health" fell under the "genuine relationships" or "sense of purpose" constructs. Additionally, items written to reflect the "sense of purpose" construct emerged as a new factor, which authors labeled "engaged citizenship". Engaged citizenship was subsequently defined as "the sense of belonging and responsibility to a larger community that promotes altruistic behavior through leadership and civic action". Engaged citizenship is culturally authentic to veterans, many of whom seek and value opportunities for community service and leadership during their transition from military to civilian life. Veteran and civilian ELS factors were identical, except for one sleep-related item, which loaded onto physical health for the mostly female civilian sample, and mental health for the mostly male veteran sample. Civilians scored higher on every subscale of the ELS and total score than veterans, with small to medium effect size differences. In the veteran sample, veterans with combat experience and service-related injuries scored lower on the ELS than veterans without combat experience or service related injuries. As the ELS was preliminarily validated in a sample of Team RWB members, the inherent bias was that members may have already been exposed to life enriching activities via participation in the organization, although the preliminary study was not designed to serve as a program evaluation framework for Team RWB. In the current study, we tested the ELS factor structure in a sample of non-Team RWB members to potentially increase generalizability to other populations of veterans; we were uncertain if veteran Team RWB members shared an inherent bias that our methods were not sensitive enough to detect when they self-selected into a fitness and social activity focused organization. Additionally, while the development and implementation of the ELS is to measure an enriched life in veterans and civilians, we limited the current study to veterans as it was the most highly prioritized need for Team RWB as veterans are an understudied population and should thus be preferred.

The goal of the current study was to use confirmatory factor analysis to validate the structure of the ELS in a sample of veterans self-identifying as not affiliated with Team RWB. Our second objective was to explore convergent validity with the Military to Civilian Questionnaire (M2C-Q), a psychometric measure of reintegration difficulties in veterans (Sayer et al., 2011). We hypothesized that as veteran participants reported higher levels of enrichment, they would report lower levels of reintegration difficulties.

MATERIALS AND METHODS

Participants

Participants were recruited electronically via direct email, partner Twitter and Facebook solicitations, and snowball sampling between March 2017 and March 2018 for a multi-purpose study. After providing informed consent, participants were directed to a secure link. Respondents self-identified as being a veteran, active duty military service members, or having no military service experience (civilians). A week after the original email was circulated, a reminder was sent to participants. A total of 1,900 respondents agreed to participate in the study through the recruitment period. Participants who were retained as part of this analysis were military veterans who self-reported that they were not members of Team RWB. We removed 800 participants from the analysis who reported that they were members of Team RWB and 78 participants who did not indicate whether they were members or not. This procedure was used in order to isolate the confirmatory factor analysis to non-Team RWB members, whom we hypothesized, may have already received life-enriching activities, based upon their exposure to Team RWB activities at the time of recruitment for the exploratory factor analysis (EFA) study (Angel et al., 2018b). Since the EFA was conducted on a sample of Team RWB members, the CFA was limited to non-Team RWB members in order to avoid overly optimistic model fit. In addition, 96 participants who started the survey but did not complete the ELS portion of the survey were removed from the analysis. Out of the remaining 926 participants, only U.S. military veterans were retained, resulting in a final sample size of 529 veterans for the CFA analysis. Each observation had complete data.

The study protocol was reviewed and approved by the Institutional Review Board at Syracuse University. Participants were informed that the purpose of the study was to develop a new instrument to track health, relationships, and sense of purpose. The average time to complete the entire 110-question survey, inclusive of the 40-item ELS, demographic variables, and other variables of interest to Team RWB, was 24 min. Qualtrics estimated that the 40-item ELS would take 8–9 min to complete by itself. No financial compensation was provided for completing the survey.

Measures

The ELS (Team Red White Blue, 2017) is a 40-item measure that assesses "enrichment," defined as physical health (having consistent physical activity, with appropriate restful sleep, nutrition, healthy weight maintenance, strength, and mobility to accomplish activities of daily living with ease); mental health (anxiety and depressive symptoms within normal limits to include controlled anger and an ability to focus, make decisions, and remember things); genuine relationships (a combination of weak and strong social ties that include close, "best-friend" types of relationships as well as a broader supportive network to provide emotional support, information, and resources); a sense of purpose (individual and shared goal driven activities integrated with positive emotion (optimism, gratitude, self-compassion, pride, open-mindedness) and positive role identity); and engaged citizenship (the sense of belonging and responsibility to a larger community that promotes altruistic behavior through leadership and civic action). ELS subscale length and example items of each scale are as follows: "genuine relationships" (11 items), "I have people in my life that are not my relatives but feel like family"; "sense of purpose" (12 items), "I have a sense of direction in my life"; "engaged citizenship" (6 items), "I feel like a leader in my community"; "mental health" (6 items), "Even when I feel nervous, anxious, or irritable, I am able to carry out day-to-day activities and responsibilities in my work and relationships,"; and "Physical Health" (5 items), "I have the strength and mobility to do all the things I need to do routinely in my life with ease". With the exception of one four-point Likert scale (i.e., item #36) that assesses the frequency, duration, and intensity of physical exercise, all items were rated on a five-point scale in increments of 25 points (ranging from zero to 100), where higher scores indicated greater enrichment.

The Military to Civilian Questionnaire (M2C-Q) (Sayer et al., 2011) is a publicly available 16-item measure that assesses veterans' post-deployment community reintegration difficulties. Areas assessed include (a) interpersonal relationships with family, friends, and peers; (b) productivity at work, in school, or at home, (c) community participation; (d) self-care; (e) leisure, and (f) perceived meaning in life. Items are rated on a 5-point Likert scale with these response options: 0 = No difficulty, 1 = A little difficulty, 2 = Some difficulty, 3 = A lot of difficulty, and 4 = Extreme difficulty. Respondents can indicate "Does not apply" for the four items that assess relationship with spouse/partner, relationship with child/children, work, and school functioning. The measure was validated in a study of 745 Iraq and Afghanistan veterans who sought medical care from the U.S. Department of Veterans Affairs (Sayer et al., 2011). The instrument was selected for convergent validity in military veterans, as we expected ELS subscales (ranging zero to 100) to be inversely related to the M2C-Q score on the basis that the ELS measures reintegration enrichment and the M2C-Q measures reintegration challenges.

Statistical Analyses

For the confirmatory factor analysis of the ELS, three models were estimated via maximum likelihood estimation with robust Huber-White standard errors (MLR) (Li, 2018), as the assumptions for standard maximum likelihood estimation (i.e., multivariate normality) were not met. Based upon the findings of the EFA (Angel et al., 2018b), a one-factor model, where all 40-items of the ELS were arranged within one latent factor, was estimated first. We then tested a four-factor model, where "sense of purpose" and "engaged citizenship" were collapsed. Then, a five-factor model was estimated, including the following constructs and items: "genuine relationships" (GR, items 1–11); "sense of purpose" (SP, items 12–23); "engaged citizenship" (EC, items 24–29); "mental health" (MH, items 30–34); and "physical health" (PH, items 35–40). The three models were compared by examining the proportion of variance accounted for, the rotated loading patterns, and the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), where smaller values indicated better fit (Burnham and Anderson, 2004). The key model fit statistics for the one-factor, four-factor, and five-factor models are shown in **Table 2**. Consistent with the findings of our EFA (Angel et al., 2018b), the five-factor model resulted in being the best fit. Residual correlations between items within the same construct were added iteratively to the five-factor model based on modification indices to improve model fit (Kline, 1998). This approach, described by Sorbom (1989) as the *post hoc* model modification approach or *post hoc* method theory, allows researchers to identify areas of theoretical misspecification within confirmatory factor analysis models, make adjustments to the theoretical model via consideration of modification indices, and generate more robust models. While there is some debate about the utility of this approach (Pan et al., 2017), we specified correlated residuals on within-construct items with modification indices greater than five (Segers, 1997) until satisfactory model fit was achieved (Sass, 2011). Residual correlations were also added to the following items due to item wording effects, such as parallel or negative wording, or item context, such as questions which reference a similar context (Schreiber et al., 2010; Asparouhov et al., 2015): GR 2 and 3; GR 3 and 10; SP 16 and 17; SP 19 and 20; SP 12 and 13; SP 18 and 20; SP 14 and 15; EC 28 and 29; EC 26 and 27; EC 25 and 28; MH 31 and 32.

Several indicators of model fit were used: the model Chi-square statistic, the Root Mean Square Error of Approximation (RMSEA), the Comparison fit index (CFI), the Tucker-Lewis fit index (TLI), and the Standardized Root Mean Square Residual (SRMR). Values of RMSEA ≤ 0.06, CFI/TLI ≥ 0.90, SRMR ≤ 0.10, and a p-value for the $\chi^2 < 0.05$ are often considered as indicating acceptable fit (Hu and Bentler, 1999; Mehmetoglu and Jakobsen, 2016). Convergent validity for the subscales was assessed by (a) estimating composite reliability (CR) for each factor, where a CR value >0.70 was considered evidence of convergent validity (Hair et al., 2008; Thornton et al., 2014), (b) examining factor loadings for statistical significance at an alpha level of 0.05 (Cole, 1987), and (c) by correlating factor scores from the validated five-factor ELS model with factor scores from the M2C-Q (Sayer et al., 2011), a measure that is theoretically inversely related to the ELS. For the purposes of standardizing comparisons between mean scores and standard deviations between the M2C-Q and the ELS, we recoded the M2C-Q response scale to correspond to the ELS (0, 25, 50, 75, 100). Discriminant validity within the five-factor ELS was assessed by calculating heterotrait-monotrait ratios of correlations (HTMT) among the five factors (subscales), using a criterion of <0.85 to indicate discriminant validity (Henseler et al., 2015). According to Henseler et al. (2015), the HTMT for two constructs is the average of the heterotrait-heteromethod correlations relative to the average of the monotrait-heteromethod correlations, as derived from the classic multitrait-multimethod matrix. We also assessed internal consistency reliability with Cronbach's alpha and adopted a criterion of >0.70 to indicate reliability (Nunnaly, 1978). All analyses were conducted with the "lavaan" (Rosseel et al., 2018) and "semTools" (Jorgensen et al., 2018) packages within the R project for statistical computing (R Core Team, 2019).

RESULTS

Participant Characteristics

Table 1 displays demographic characteristics for the sample which included a total of 529 veterans. Over 78% of veterans in our sample were male, and 60% of the sample was between the ages of 40 and 60, while 30% of veterans were younger than 40. Almost 80% of the sample was married or in a partnership and

TABLE 1 | Demographic characteristics of the study sample (n = 529).

	n	%
Gender		
Male	416	78.60
Female	113	21.40
Age		
20 to 40	167	31.60
41 to 60	321	60.70
61 to 80	32	6.00
Refused	9	1.70
Race/Ethnicity		
American Indian/Alaskan Native	5	0.90
Asian/Pacific Islander/Native Hawaiian	21	4.00
Black/African American	89	16.80
Hispanic	44	8.30
White	355	67.10
Other	13	2.50
Refused	2	0.40
Employment Status		
Employed	334	63.14
Unemployed	63	11.91
Retired	35	6.62
Student	46	8.67
Disabled	12	2.28
Other	39	7.38
Marital Status		
Single	54	10.21
Married/Partnership	420	79.39
Divorced/Separated	50	9.45
Widowed	2	0.38
Refused	3	0.57
Educational Attainment		
High School Degree	6	1.13
Some College	71	13.42
Associate's Degree	34	6.43
Bachelor's Degree	146	27.60
Graduate School	265	50.09
Other/Refused	7	1.33
Annual Income (USD)		
0 to 24,999	27	5.10
25,000 to 49,999	53	10.02
50,000 to 74,999	75	14.18
75,000 to 99,999	94	17.77
100,000+	273	51.61
Refused	7	1.32
Military Branch[a]		
Army/Army Reserve/Army National Guard	338	63.89
Navy/Navy Reserve	137	25.89
Air Force/Air Force Reserve/Air National Guard	111	20.98
Marine Corps/Marine Corps Reserve	80	15.12
Coast Guard/Coast Guard Reserve	14	2.64
Combat experience, yes	384	72.7
Service-related injury, yes	351	66.6

[a] *Participants could select multiple response options for this question.*

over 77% of the veterans had at least an undergraduate college education. Sixty-three percent of veterans were employed, while 11% were unemployed.

Sixty-four percent of veterans had served in the Army, Army National Guard or Army Reserve. Seventy-three percent of veterans in the sample had combat experience, and 66.6% said they had a service-related injury.

Model Fit Statistics

In **Table 2**, we provide the model-fit statistics for the one-factor, four-factor, and five-factor ELS models. Results showed that the five-factor model was a good fit to the data according to the RMSEA, CFI, TFI, and SRMR statistics, while the one-factor and four-factor models indicated inadequate fit to the data. In addition, since the AIC and BIC values were lower for the five-factor model (AIC = 182,458.85, BIC = 182,890.22) than the one-factor model (AIC = 185,958.80, BIC = 186,300.48) and the four-factor model (AIC = 183,586.04, BIC = 183,953.34), the five-factor ELS model should be preferred. This model is shown in **Figure 1**. We computed average variance extracted for each latent construct in order to determine the amount of variance explained within each construct by its items and obtained the following results: sense of purpose = 0.55, genuine relationships = 0.46, engaged citizenship = 0.54, mental health = 0.61, physical health = 0.43. Malhotra and Dash (2011) indicated that average variance extracted is "a more conservative measure than CR. On the basis of CR alone, the researcher may conclude that the convergent validity of the construct is adequate, even though more than 50% of the variance is due to error" (p. 702). A full list of the items are described in Angel et al. (2018b) and available from Team Red White Blue (2017).

Internal Consistency Reliability

We assessed internal consistency reliability with Cronbach's alpha and adopted a criterion of >0.70 to indicate reliability (Nunnaly, 1978). Results showed that the five factors of the ELS exhibited satisfactory internal consistency reliability: genuine relationships, $\alpha = 0.90$; sense of purpose, $\alpha = 0.93$; engaged citizenship, $\alpha = 0.89$; mental health, $\alpha = 0.88$; and physical health, $\alpha = 0.78$.

TABLE 2 | Model fit statistics for the 40-item Enriched Life Scale (ELS) (*n* = 529).

Model	χ^2	*df*	*p*-value	RMSEA	CFI	TLI	SRMR
One-factor model	4,516.42	740	<0.01	0.10	0.64	0.62	0.10
Four-factor model	2,648.67	734	<0.01	0.07	0.82	0.81	0.07
Five-factor model	1,731.51	719	<0.01	0.05	0.90	0.90	0.06

df, degree of freedom; RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker-Lewis index; SRMR, standardized root mean square residual. The following residual correlations were added to the five-factor ELS model: GR 2 and 3; GR 3 and 10; SP 16 and 17; SP 19 and 20; SP 12 and 13; SP 18 and 20; SP 14 and 15; EC 28 and 29; EC 26 and 27; EC 25 and 28; MH 31 and 32. The five-factor model without residual correlations exhibited the following fit statistics: χ^2 (730) = 2,317.55, $p < 0.01$, RMSEA = 0.06, CFI = 0.85, TLI = 0.84, SRMR = 0.07, AIC = 183,176.83, BIC = 183,561.22.

Convergent Validity and Reliability

Standardized factor loadings for the five-factor ELS model are shown in **Figure 1**. Results showed that all unconstrained factor loadings within the five factors were statistically significant at an alpha level of 0.05. Furthermore, composite reliability (CR) indices for each factor, which indicated whether items within the same factor measured the same construct (Hair et al., 2008; Thornton et al., 2014), were >0.70: genuine relationships, CR = 0.91; sense of purpose, CR = 0.93; engaged citizenship, CR = 0.89; mental health, CR = 0.88; and physical health, CR = 0.80. Given the criteria outlined in Cole (1987), Hair et al. (2008), and Thornton et al. (2014), results showed that the five factors within the ELS demonstrated convergent validity. We also assessed convergent validity of the five-factor model by calculating Pearson correlation coefficients between scores from the validated five-factor ELS model and factor scores from the M2C-Q, which we hypothesized would be inversely related. Given that all Pearson correlation coefficients were negative and exhibited p-values < 0.05, further evidence of convergent validity for the ELS was provided.

Discriminant Validity

We examined discriminant validity within the five-factor ELS by calculating HTMT among the five factors, using a criterion of <0.85 to indicate discriminant validity (Henseler et al., 2015). Results showed that the HTMT ratios between each of the five factors in the ELS were less than 0.85, providing initial evidence of discriminant validity within the ELS (**Table 3**). Mean and standard deviations for the final ELS scales ranged from 55.71 (SD = 19.52) for physical health to 75.51 (SD = 16.93) for genuine relationships (**Table 3**). On a scale of 0 to 100, the mean score on the M2C-Q was 25.43 (SD = 20.51).

DISCUSSION

The first aim of this study was to confirm the factor structure of the ELS in veterans not affiliated with Team RWB. Our second goal was to determine if the ELS would have convergent validity with the Military to Civilian Questionnaire, a psychometric measure of reintegration difficulties experienced by veterans.

The results of the CFA indicated that the hypothesized five-factor structure was the most adequate for the ELS, and all items contributed significantly to their corresponding factor: genuine relationships, sense of purpose, engaged citizenship, physical health, and mental health. The model-based reliability for each construct was also excellent. Per the HTMT ratios, the constructs within the ELS were different enough to demonstrate internal discriminant validity.

This finding in a non-Team RWB sample reinforces our initial conceptualization describing veteran transition and as having physical and mental health, people, purpose, and the newly emerged engaged citizenship (continued service) construct as foundational tenets of an enriched life (Angel et al., 2018b). Mean scores and standard deviations were also comparable to those previously reported in the Team RWB sample (Angel et al., 2018b).

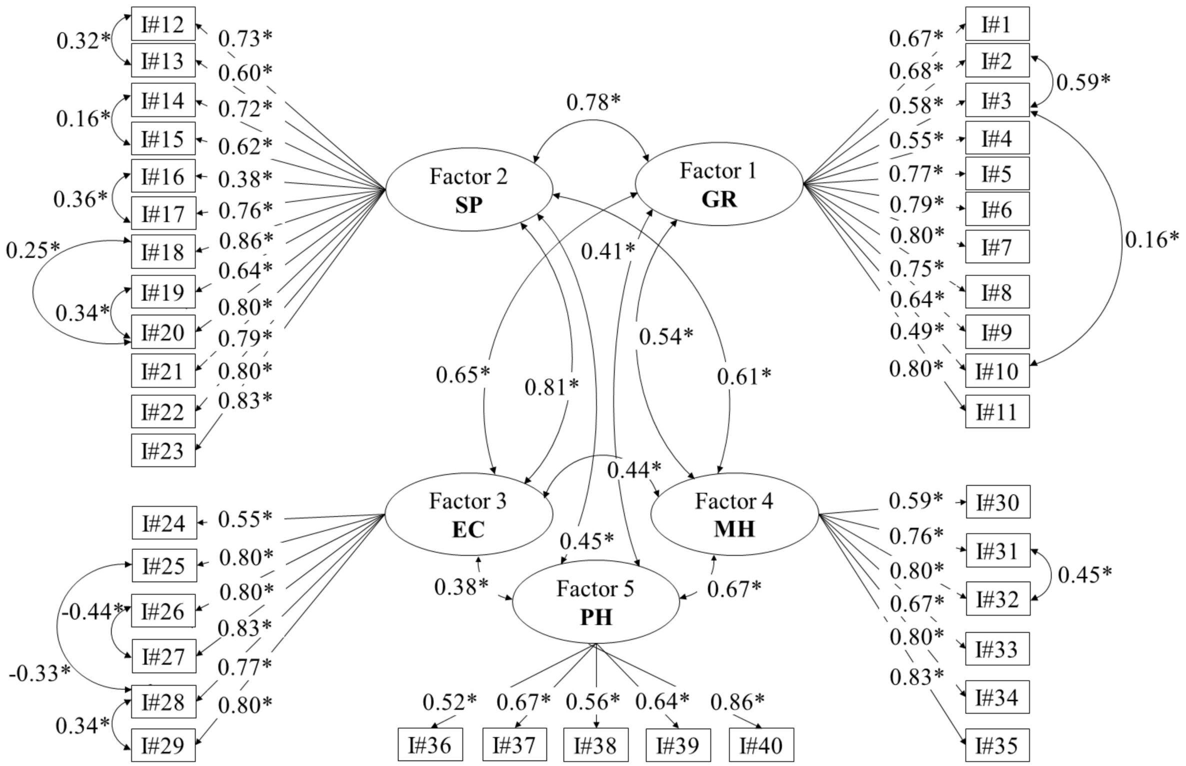


FIGURE 1 | Five-factor model of the Enriched Life Scale with factor loadings and factor covariances. Asterisks indicate statistical significance at $p < 0.05$. SP = sense of purpose, EC = engaged citizenship, PH = physical health, MH = mental health, GR = genuine relationships.

TABLE 3 | Construct validity results for the five-factor ELS model (n = 529).

		Pearson *r*	**Heterotrait-Monotrait Ratios of Correlations (Pearson *r* Correlations)**				
	M^a (SD)	**M2C-Q^b**	**GR**	**SP**	**EC**	**MH**	**PH**
GR	75.51 (16.93)	–0.59*	–				
SP	74.92 (17.12)	–0.69*	0.79 (0.72*)	–			
EC	61.52 (21.53)	–0.55*	0.70 (0.62*)	0.80 (0.72*)	–		
MH	70.96 (19.38)	–0.77*	0.51 (0.45*)	0.63 (0.57*)	0.42 (0.38*)	–	
PH	55.71 (19.52)	–0.48*	0.40 (0.33*)	0.47 (0.40*)	0.38 (0.32*)	0.68 (0.57*)	–
ELS Total Score^c	67.62 (15.17)	–0.78*	–	–	–	–	–

**Pearson r coefficient p-value < 0.05. ^{a}Means and standard deviations for average construct scores.^{b}M2C-Q Score Mean = 25.43 (SD = 20.51), RMSEA = 0.10, SRMR = 0.07, CFI = 0.82, TFI = 0.80, Chi-Square = 1,746.43, Chi-Square p-value < 0.001, α = 0.92. ^{c}The total Enriched Life Scale (ELS) score is calculated by taking the mean of the five scales (GR, SP, EC, MH, and PH).*

The study had several limitations which are noted here, and could be addressed in subsequent research projects. Our sampling approach, which sought to recruit participants via a general request for participation across social media channels and targeted email by partner organizations, may have influenced participation. Based upon the broad solicitation for participation over the course of a year, we cannot tell how many individuals were exposed to a request for participation nor the number of potential respondents the study might have had if all potential respondents had consented to participation. Nevertheless, an achieved sample of over 500 veteran respondents is considered very good for understanding the relationship between latent factors and their constructs, which was the primary goal of the study.

Another potential bias was that we did not assess for the multitude of ways that participants might have been involved in life enriching activities. Based upon the survey recruitment pools, respondents are likely to come from a variety of veteran enriching programs, although we specifically excluded members who self-identified as Team RWB members. Social desirability bias might have influenced the study findings and based upon the recruitment methodology of anonymous participants, we were unable to track them over time. Longitudinal tracking of potential changes in ELS scores is another area of future research. Additionally, our analysis showed that four out of five model fit indices calculated in this study for the M2C-Q indicated poor fit with a one-factor solution. While it was beyond the scope of this paper to investigate an alternative factor structure for the M2C-Q, future studies should consider examining a multi-factor structure for the M2C-Q.

Another limitation of the study is the gender imbalance of participants. While the majority of veteran participants were men, 20% of our study participants were veteran women, twice the number of women veterans comprising the total veteran population in the United States as of 2015 (Office of Data Governance and Analytics, 2017). Understanding the factor structure of ELS for women veterans specifically is an important area for future research. We only tested convergent validity of the ELS with one other measure, which has been done in other studies, such as ones with the PHQ-9 (Cameron et al., 2008) and the SF-6D (Kontodimopoulos et al., 2009). However, future studies should consider testing the convergent validity of the ELS with other multidimensional measures. In addition, as all the scales in the study were self-reported, construct validity of the ELS should be evaluated using different methods in future research, including other types of reporting and additional behavioral measures.

Our results supported our hypothesis that each of the five ELS factors (and ELS total score) were negatively associated with the M2C-Q questionnaire, indicating that veterans who experience greater physical health, mental health, genuine relationships, sense of purpose, and engaged citizenship report fewer reintegration difficulties. This finding has important implications for the implementation of the ELS as a practical assessment tool for veteran health and wellbeing and how it can be integrated into the broader portfolio of clinical assessment tools. The M2C-Q focuses on reintegration challenges. Unlike the battery of other available psychiatric diagnostic and substance misuse instruments administered by the Veterans Health Administration (VHA) and Department of Defense (Patient Health Questionnaire-2, Patient Health Questionnaire-9, Primary Care Post-Traumatic Stress Disorder screen, Alcohol Use Disorders Identification Test-Consumption, Post-Deployment Health Assessment) (Panaite et al., 2018), the M2C-Q is used for screening transition stress related to community integration, personal relationships, self-care, and meaning in life. Consequently, the M2C-Q provides insight into transition-related problems that are neither diagnostic nor reflective of specific mental health related pathology.

While we have very limited visibility of screening instruments implemented in VHA clinical sites and other leading health institutions serving veterans, what we can determine based upon review of publicly available websites via Google search (which may be the only information available to veterans and the layperson community), is that currently veterans seeking information from the VHA website are offered four mental health screening assessments (PTSD screening via the PTSD Check List (PCL); depression screening via the Patient Health Questionnaire-9 (PHQ-9); substance abuse screening via the Alcohol, Smoking and Substance Involvement Screening Test (ASSIST); and alcohol use screening via the Alcohol Use Disorders Identification Test for Consumption (AUDIT-C)[2]. While critical to directing veterans to mental health resources and potentially a starting point for much needed mental health intervention, arguably greater emphasis could potentially be placed on illuminating a broader spectrum of mental health. The ELS's focus on "what goes right in life," coupled with the existing strain focused assessments, could help to reframe health assessment aligned to a comprehensive wellness framework, where the underlying message delivered to veteran respondents communicates an expectation of thriving, along with assessment of potential challenges. Doing so potentially helps derail the "victimhood" narrative (Kleykamp and Hipes, 2015), which is perpetuated when health care institutions remain focused on a paradigm of expected brokenness.

By current assessment standards, it is not possible to tell which veterans screen positive for post-traumatic stress, but nevertheless are still leading a life that they feel is filled with purpose, direction, and shared goals with others. The lived experience of veterans demonstrates that not only can both pathways be possible, but we recommend that communicating both as part of an overall health status is critical, if clinicians are to keep veterans' holistic health needs at the center of their wellness journeys home. The VHA is leading in so far that they are making strides through the development of their "Whole Health for Life" platform, and the development of their Personal Health Inventory[3], yet these advances have yet to translate into a publicly available, screening instrument for veterans. The ELS could assist in making that possible.

The ELS has demonstrated tremendous promise for use as a general wellness assessment tool in the civilian community as well. In our preliminary study documenting the ELS's factor structure, both veteran and civilian versions were nearly identical, with only one item related to sleep falling on the civilian physical health scale and the veteran mental health scale. Our next research steps will be to confirm the factor structure in a sample of civilian community members. We are encouraged about growing evidence that the ELS is a measure of well-being for all people.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of The Institutional Review Board at Syracuse

[2] https://www.myhealth.va.gov/mhv-portal-web/screening-tools

[3] https://www.va.gov/PATIENTCENTEREDCARE/resources/personal-health-inventory.asp

University with electronic informed consent from all subjects. All subjects gave electronic informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Institutional Review Board at Syracuse University.

AUTHOR CONTRIBUTIONS

CA, NA, BY, JP, and MW designed the ELS, conceived the study, and provided conceptual guidance and commentary. CA and RL collected the data. MW and JM analyzed and interpreted the data. CA, MW, RL, and JM contributed to writing the manuscript. All authors reviewed and approved the final version for publication.

ACKNOWLEDGMENTS

We wish to thank Team Red, White & Blue, the Institute for Veterans and Military Families at Syracuse University, Michael D. Boll of the New Jersey Veterans Network, and William D. Walsh of Walsh Public Safety Consulting and Training for assisting us with the recruitment of participants. Investigators interested in using the ELS should contact the first author (caroline.angel@teamrwb.org).

REFERENCES

Angel, C. M. (2016). Resilience, post-traumatic stress, and posttraumatic growth: veterans' and active duty military members' coping trajectories following traumatic event exposure. *Nurse Educ. Today* 47, 57–60. doi: 10.1016/j.nedt.2016.04.001

Angel, C. M., and Armstrong, N. J. (2016). *Enriching Veterans Lives Though an Evidence Based Approach: A Case Illustration of Team Red, White & Blue (Measurement and Evaluation Series, Paper 1)*. Syracuse, NY: Institute for Veterans and Military Families, Syracuse University.

Angel, C. M., Smith, B. P., Pinter, J. M., Young, B. B., Armstrong, N. J., Quinn, J. P., et al. (2018a). Team Red, White & Blue: a community-based model for harnessing positive social networks to enhance enrichment outcomes in military veterans reintegrating to civilian life. *Transl. Behav. Med.* 8, 554–564. doi: 10.1093/tbm/iby050

Angel, C. M., Woldetsadik, M. A., Armstrong, N. J., Young, B. B., Linsner, R. K., Maury, R. V., et al. (2018b). The enriched life scale: development, exploratory factor analysis, and preliminary construct validity for US military veterans and civilian samples. *Transl. Behav. Med.* 8, 554–564. doi: 10.1093/tbm/iby109

Asparouhov, T., Muthen, B., and Morin, A. J. S. (2015). Bayesian structural equation modeling with cross-loadings and residual covariances: comments on Stromeyer et al. *J. Manag.* 41, 1561–1577. doi: 10.1177/0149206315591075

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Soc. Methods Res.* 33, 261–304. doi: 10.1177/0049124104268644

Butler, J., and Kern, M. L. (2016). The PERMA_Profiler: a brief multidimensional measure of flourishing. *Int. J. Wellbeing* 6, 1–48. doi: 10.5502/ijw.v613.526

Cameron, I. M., Crawford, J. R., Lawton, K., and Reid, I. C. (2008). Psychometric comparison of PHQ-9 and HADS for measuring depression severity in primary care. *Br. J. Gen. Pract.* 58, 32–36. doi: 10.3399/bjgp08x263794

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *J. Consult. Clin. Psychol.* 55, 584–594. doi: 10.1037/0022-006X.55.4.584

Cornum, R., Matthews, M. D., and Seligman, M. E. P. (2011). Comprehensive soldier fitness: building resilience in a challenging institutional context. *Am. Psychol.* 66, 4–9. doi: 10.1037/a0021420

Diener, E. D., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess* 49, 71–75.

Diener, E. D., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., et al. (2010). New well-being measures: short scales to assess flourishing and positive and negative feelings. *Soc. Indic. Res.* 97, 143–156. doi: 10.1007/s11205-009-9493-y

Elnitsky, C. A., Fisher, M. P., and Blevins, C. L. (2017). Military service member and veteran reintegration: a conceptual analysis, unified definition, and key domains. *Front. Psychol.* 8:369. doi: 10.3389/fpsyg.2017.00369

Green, K. T., Hayward, L. C., Williams, A. M., Dennis, P. A., Bryan, B. C., Taber, K. H., et al. (2014). Examining the factor structure of the Connor-Davidson Resilience Scale (CD-RISC) in a post-9/11 U.S. military veteran sample. *Assessment* 21, 443–451. doi: 10.1177/1073191114524014

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2008). *Multivariate Data Analysis*, 7th Edn. New Jersey, NJ: Pearson Education.

Henseler, J., Ringle, C. M., and Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Mark. Sci.* 43, 115–135. doi: 10.1007/s11747-014-0403-8

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Modeling* 6, 1–55. doi: 10.1080/10705519909540118

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A., Rosseel, Y., Miller, P., Quick, C., et al. (2018). *Package 'semTools'*. Available at: https://cran.r-project.org/web/packages/semTools/semTools.pdf (accessed November 20, 2018).

Kleykamp, M., and Hipes, C. (2015). Coverage of veterans of the wars in Iraq and Afghanistan in the US media. *Soc. Forum* 30, 348–368. doi: 10.1111/socf.12166

Kline, R. B. (1998). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guildford Press.

Kontodimopoulos, N., Pappa, E., Papadopoulos, A. A., Tountas, Y., and Niakas, D. (2009). Comparing SF-6D and EQ-5D utilities across groups differing in health status. *Qual. Life Res.* 18, 87–97. doi: 10.1007/s11136-008-9420-8

Li, J. (2018). Probability of superiority SEM (PS-SEM): detecting probability based multivariate relationships in behavioral research. *Front. Psychol.* 9:883. doi: 10.3389/fpsyg.2018.00883

Linton, M.-J., Dieppe, P., and Medina-Lara, A. (2018). Review of 99 self- report measures for assessing well-being in adults: exploring dimensions of well-being and developments over time. *BMJ Open.* 6:e010641. doi: 10.1136/bmjopen-2015-010641

Malhotra, N. K., and Dash, S. (2011). *Marketing Research an Applied Orientation*. London: Pearson Publishing.

Mehmetoglu, M., and Jakobsen, T. G. (2016). *Applied Statistics Using Stata: A Guide for the Social Sciences*. Thousand Oaks, CA: SAGE.

Mobbs, M. C., and Bonanno, G. A. (2018). Beyond war and PTSD: the crucial role of transition stress in the lives of military veterans. *Clin. Psychol. Rev.* 59, 137–144. doi: 10.1016/j.cpr.2017.11.007

Nunnaly, J. (1978). *Psychometric Theory*. New York, NY: McGraw-Hill.

Office of Data Governance and Analytics (2017). Women Veterans Report. Available at: https://www.va.gov/vetdata/docs/SpecialReports/Women_Veterans_2015_Final.pdf (accessed September 30, 2019).

Panaite, V., Brown, R., Henry, M., Garcia, A., Powell-Cope, G., Vanderploeg, R. D., et al. (2018). Post-deployment mental health screening: a systematic review of current evidence and future directions. *Adm. Policy. Ment. Health.* 45, 850–875. doi: 10.1007/s10488-018-0869-7

Pan, J., Ip, E. H., and Dube, L. (2017). An alternative to post-hoc model modification in confirmatory factor analysis: the bayesian lasso. *Psychol. Methods* 22, 687–704. doi: 10.1037/met0000112

Pavot, W., and Diener, E. (2009). *Review of the Satisfaction With Life Scale. In Assessing well-Being*. Dordrecht: Springer, 101–117.

R Core Team (2019). *R: A Language and Environment for Statistical computing*. Available at: https://www.r-project.org/ (accessed November 20, 2018).

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., et al. (2018). *Package 'Lavaan'*. Available at: https://cran.r-project.org/web/packages/lavaan/lavaan.pdf (accessed November 20, 2018).

Ryan, R. M., and Deci, E. L. (2001). On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annu. Rev. Psychol.* 52, 141–166. doi: 10.1146/annurev.psych.52.1.141

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *J. Pers. Soc. Psychol.* 57, 1069–1081. doi: 10.1037/0022-3514.57.6.1069

Ryff, C. D. (2018). Well-being with soul: science in pursuit of human potential. *Perspect. Psychol. Sci.* 13, 242–248. doi: 10.1177/1745691617699836

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *J. Psychoeduc. Assess.* 29, 347–363. doi: 10.1177/0734282911406661

Sayer, N. A., Frazier, P., Orazem, R. J., Murdoch, M., Gravely, A., Carlson, K. F., et al. (2011). Military to civilian questionnaire: a measure of postdeployment community reintegration difficulty among veterans using department of veterans affairs medical care. *J. Trauma Stress* 24, 660–670. doi: 10.1002/jts.20706

Schreiber, J. B., Nora, A., Stage, F. K., Barrow, E. A., and King, J. (2010). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338.

Segers, A. H. (1997). Assessing the unidimensionality of measurement: a paradigm and illustration within the context of information systems. *Omega* 25, 107–121. doi: 10.1016/s0305-0483(96)00051-5

Seligman, M. E. P. (2011). *Flourish: A Visionary new Understanding of Happiness and Well-Being*. New York, NY: Free Press.

Seligman, M. E. P., and Csikszentmihalyi, M. (2000). Positive psychology: an introduction. *Am. Psychol.* 60, 410–421.

Sorbom, D. (1989). Model modification. *Psychometrika* 54, 371–384.

Team Red White Blue, (2017). *The Enriched Life Scale*. Tampa: Team Red, White & Blue.

Thornton, S. C., Henneberg, S. C., and Naude, P. (2014). Conceptualizing and validating organizational networking as a second-order formative construct. *Ind. Market. Manag.* 43, 951–966. doi: 10.1016/j.indmarman.2014.05.001

Umucu, E., Wu, J., Sanchez, J., Brooks, J. M., Chiu, C., Tu, W., et al. (2019). Psychometric validation of the PERMA-profiler as well-being measure for student veterans. *J. Am. Coll. Health* doi: 10.1080/07448481.2018.1546182 [Epub ahead of print].

VanderWeele, T. J. (2017). On the promotion of human flourishing. *Proc. Natl. Acad. Sci. U.S.A.* 31, 8148–8156.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 October 2019
doi: 10.3389/fpsyg.2019.02423

Measuring the Psychological Security of Urban Residents: Construction and Validation of a New Scale

Jiaqi Wang[1], *Ruyin Long*[1,2]*, *Hong Chen*[1]* and *Qianwen Li*[1]

[1] School of Management, China University of Mining and Technology, Xuzhou, China, [2] Research Center for Energy Economics, School of Business Administration, Henan Polytechnic University, Jiaozuo, China

With the acceleration of urbanization in developing countries, resources relating to medical care and the environment are becoming increasingly scarce, and the negative spillover effects brought about by scientific and technological progress have also significantly increased the pressure on urban residents. The psychological security of urban residents has recently undergone significant change. This paper introduces psychological security into the area of urban residents' lives, defines the concept of urban residents' psychological security, and presents the development and validation of the Urban Residents Psychological Security Scale (URPS). By considering psychological indicators, this paper supplements our knowledge on environmental indicators such as the risk perception of environmental pollution and climate change, and social indicators such as urban belongingness and the risk perception of technology which verifies the negative spillover effects of technological development. Based on a literature search and consideration of grounded theory (25 urban residents' in-depth interview records), the psychological security of urban residents is divided into three dimensions: self-psychological security, social environmental security, and natural environmental security, consisting of 20 items. In this study, 802 questionnaires were completed by participants. We determined that the URPS scale has good reliability and validity using exploratory factor analysis and confirmatory factor analysis, and conclude that the scale can be used as an effective measurement tool for urban residents' psychological security. The development of this scale has important theoretical and practical significance in helping city managers better understand the residents' demands and to monitor the implementation effects of policies.

Keywords: psychological security, urban residents, scale development, grounded theory, quantitative analysis

OPEN ACCESS

Edited by:
Elisa Pedroli,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Jingyang Zhou,
Shandong Jianzhu University, China
Dan O'Brien,
Northeastern University, United States

****Correspondence:***
Ruyin Long
longruyin@163.com
Hong Chen
hongchenxz@163.com

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *03 April 2019*
Accepted: *11 October 2019*
Published: *25 October 2019*

Citation:
Wang J, Long R, Chen H and Li Q (2019) Measuring the Psychological Security of Urban Residents: Construction and Validation of a New Scale. Front. Psychol. 10:2423. doi: 10.3389/fpsyg.2019.02423

INTRODUCTION

With the acceleration of economic development and urbanization in developing countries, profound changes have occurred in aspects such as the economic system, social structure, and values. These changes modified people's original ways of thinking and even their lifestyles. The inherent requirements for people in relation to quality of life and environmental safety are rapidly coming to developed countries. In addition, studies increasingly show that air pollution, soil pollution, climate change, and so on, will not only affect people's physical health (Burnett et al., 2018; Zhang et al., 2018) but also indirectly or directly harm people's mental health (Evans, 2003; Chen et al., 2018; Obradovich et al., 2018). The continuous advancement of technologies such as the Internet and artificial intelligence has a significantly positive impact on remotely connecting relationships and increasing productivity but can also lead to negative effects such as unwanted personal information disclosure, Internet addiction, and social anxiety (Chesley, 2005; Gámez-Guadix and Calvete, 2016; Jia et al., 2017). Few researchers have systematically studied the negative spillover effects of technological progress, and even fewer have incorporated these effects into psychology.

As a decisive factor of mental health, psychological security has been widely concerned. Maslow defined psychological security as "a feeling of confidence, safety and freedom that separates from fear and anxiety, and especially the feeling of satisfying one's needs now (and in the future)."

In previous studies, most research on psychological security has focused on the workplace (Probst, 2002; Hu et al., 2018), and the psychological security of urban residents has not received sufficient attention. The psychological security of urban residents has mainly consisted of fear of crime, public security or social security, most of which are directly related to social factors such as public security, food safety, and medical supervision. However, insecurity is shaped by everyday experiences and often is more related to experiences of living in a risky society than to only criminal incidents (Garland, 2000). Therefore, the psychological security of urban residents should be a complex multidimensional structure rather than a simple one-dimensional structure. By analyzing and summarizing the literature, the psychological security of urban residents can be divided into three categories: psychological, social, and environmental. Most studies have focused only on the influences of individual psychological factors and social factors (Edmondson and Lei, 2014; Oishi and Kesebir, 2015; Soto, 2015), whereas insufficient attention has been paid to the effects of environmental factors. The traditional structural dimension cannot adapt to actual needs, and there is not currently available a scale that matches the actuality.

On the basis of the arguments developed above, we summarized the concept of residents' psychological security at the city level by combining them with the practical needs, explored and developed the Urban Residents' Psychological Security Scale (URPS) scale to be applicable to the current environment, and verified the applicability of the three-dimensional structure including psychological, social and environmental factors. As shown in **Figure 1**, on the basis of traditional indicators, such as interpersonal security, certainty in control, social risk perception, and occupational security, we incorporated environmental pollution risk perception and natural disaster risk perception into the structure of urban residents' psychological security, while considering the indicators of technology risk perception, urban belongingness and climate change risk perception. Among the indicators, urban belongingness was a unique indicator of psychological security at the city level.

The URPS scale developed in this paper could help city managers understand the security status of urban residents, including psychological, social, and environmental aspects, and could help relevant departments formulate targeted intervention policies. In the future, this scale is expected to effectively enhance urban attractiveness, improve the urban integration of migrants and reduce the crime rate.

The remainder of this paper is arranged as follows. Section "Literature Review" describes the related research on the psychological security of urban residents. The qualitative analysis method of Grounded Theory is used to construct the initial scale of psychological security of urban residents in section "Initial Scale Construction Based on Grounded Theory." In section "Quantitative Method," we purify the scale and test its reliability and validity using data from pre-survey and formal survey. The results of this study are discussed further in section " Discussion and Conclusion" and the conclusions are given. The section "Limitations and Future Studies" is the limitations of this study and directions for future research.

LITERATURE REVIEW

Concept and Dimensions

Cong and An (2004) defined psychological security according to Maslow (1942) as the presentiment that may arise from dangers or risks in the physiology or the psychology of the individual, as well as the sense of powerfulness and powerlessness of the individual in dealing with dangers or risks, mainly related to the sense of certainty and controllability. It is widely used by researchers (Sun and Yao, 2009; Zhao and Jing, 2013; Yu and Zhao, 2016). Hart et al. (2005) and Hart (2014) believes that psychological insecurities refer to each individual's anxiety about potential harm and threat. Obviously, the sense of psychological security is a subjective judgment of whether the individual's environment is deterministic and controllable, and the state of consciousness based on his or her own personality traits.

According to the above literature, the characteristics of psychological security can be summarized as follows: (1) psychological security is an emotional experience perceived by the individual. This emotional experience is derived from external stimuli and is determined by both the intensity of the stimulus and the psychological quality of the individual. (2) The expression of psychological security is mainly the certainty, control, and risk premonition felt by the individual. (3) Psychological security will affect physical and mental health. Individuals with higher psychological security will experience more confidence and freedom while individuals with lower psychological security are more prone to anxiety or fear, and even depression. Differences in the personality and environmental perception of individuals determine the level of the individual's trust in the outside world, and is self-centered and based on the objective environment. Individuals then further evaluate and decide whether or not the outside world is safe, and that usually connects with the degree of recognition with the outside world or the degree of willingness to contribute to it. Therefore, the connotations of individual psychological security

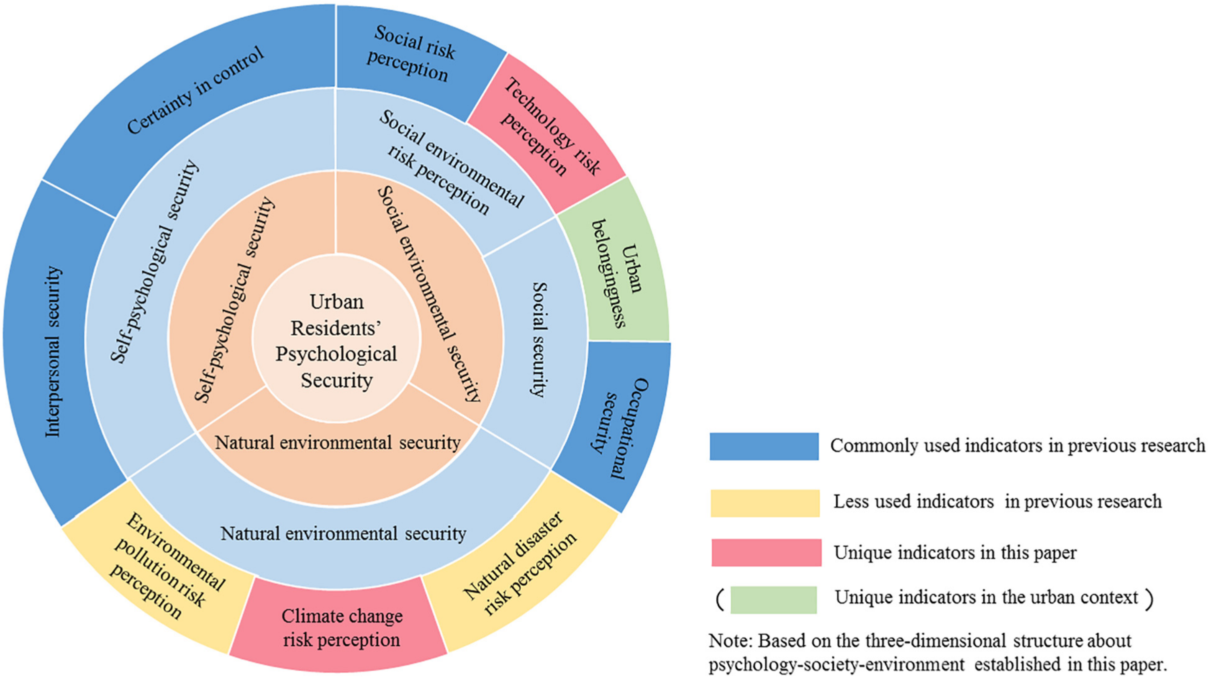

Urban Residents' Psychological Security
Self-psychological security
Social environmental security
Natural environmental security
Self-psychological security
Social environmental risk perception
Social security
Natural environmental security
Certainty in control
Social risk perception
Technology risk perception
Urban belongingness
Occupational security
Natural disaster risk perception
Climate change risk perception
Environmental pollution risk perception
Interpersonal security
Commonly used indicators in previous research
Less used indicators in previous research
Unique indicators in this paper
(Unique indicators in the urban context)
Note: Based on the three-dimensional structure about psychology-society-environment established in this paper.

FIGURE 1 | The model of urban residents' psychological security.

change with the environmental background, for example, individual psychological security in the workplace. Carmeli and Gittell (2009) effectively combined personal perceptions in the social and work fields, and believed that psychological security refers to people's views on their social environment and work environment, as well as their perceived reactions to risk-taking behaviors in the workplace. By combining individuals' perceptions of themselves, society and the urban environment, we attempted to introduce psychological security into the background of urban life, and we defined urban residents' psychological security as the risk judgment of individuals living in cities of their own urban living conditions based on past experience or intuition.

All human emotions are derived from the direct feelings of the heart. The certainty in control is one of the important and widely used dimensions of psychological security (Zhao and Jing, 2013; Yu and Zhao, 2016). Loss of control not only changes the individual's perceptions, beliefs, and behaviors but also affects their physical and mental health (Whitson and Galinsky, 2008). At the same time, individuals in the city will also have various types of interpersonal needs in their social lives. Demir (2008), Edmondson and Lei (2014), and Inoue et al. (2016) found that there is a significant correlation between interpersonal relationships and the sense of security. Safe and supportive social relationships are not only beneficial to individuals (Kagan, 2009) but also promote prosocial behavior (Mikulincer and Shaver, 2007). Negative interpersonal events can cause individuals to feel anxiety and other similar emotions while positive interpersonal experiences will effectively reduce attachment anxiety (Davila and Sargent, 2003; Zhang, 2009). Individuals with higher levels of interpersonal trust and interpersonal security will perceive fewer negative events and thus have a higher sense of psychological security.

The psychological security of residents is also affected by external objective factors. In addition to the economic development of the city, the key factors determining whether the local residents leave and whether foreigners stay for a long time are people's familiarity with the urban environment and the degree of recognition with the urban atmosphere. As well as the economic level of the city, the key factors determining whether the local residents leave and whether transient populations stay for a long time are people's familiarity with the urban area and the degree of recognition with the urban atmosphere. This emotional element is known as urban belongingness, a unique indicator of psychological security in the urban context. The individual's demand for belongingness is due to the desire for security. The need for a sense of belonging stems from the desire for security. Factors such as equity protection, housing status, and social integration will reduce the sense of belongingness and the urban identity of the non-native population who work and live in the city, which will result in their relatively isolated social relationships, cultural activities, and political participation, thus affecting the city's social and economic development. Economic factors also determine the psychological security of urban residents to a certain extent (Van Hal, 2015), which is reflected in occupational stability and occupational risk. In addition, a large number of studies have shown that the fear of crime in terms of social security factors will increase people's psychological pressure (Astell-Burt et al., 2015), and have a negative impact on their sense of security and well-being (Foster et al., 2016;

Prieto Curiel and Bishop, 2017). Carter et al. (2011), Ross and Hill (2013), and Tseng et al. (2017) have also found such negative effects from food insecurity.

In recent years, the emergence of natural disasters and environmental pollution has caused people to frequently feel a sense of having lost control. Publicity and education on energy conservation, emission reduction, and green and low carbons have made more people aware of the urgency of environmental protection issues. Doherty and Clayton (2011) found that climate change threatens the emotional health of people by making them worry or feel uncertain about future risks. The haze affects the psychological and physical health of people who live in a polluted area. The perception of smog risk even leads to the outflow of talents in smog-polluted areas (Lu and Long, 2018). Sekulova and Van den Bergh (2016) argue that natural disasters, which may be considered to be large-scale traumatic events, not only cause considerable material losses, but also can seriously impair psychological health. The challenge of tackling climate change and environmental pollution has become increasingly critical, and a series of social surveys are needed to improve the ability of psychologists and governments to cope with the relevant impacts of this.

In conclusion, the psychological security of urban residents is the risk judgment of individuals living in cities for their own state and urban living conditions based on past experience or intuition. The dimensions include (1) self-psychological security, that is, the individual's safety expectations for future life based on past life experiences, and their positive experiences of maintaining a favorable position in their own situation through the process of interpersonal interaction. (2) Social environmental security, reflecting residents' psychological attachment and identity with the city they live in, and their comprehensive risk perception of their social environment, urban atmosphere, and professional status. (3) Natural environmental security, that is, the risk perception of urban residents toward their living urban natural environment.

TABLE 1 | Psychological security dimension.

Study	Dimension	Research object
Maslow, 1942	Safety, belongingness and receiving love and affection	–
Brown and Leigh, 1996	Supportive management, role clarity, contribution, recognition, self-expression and challenge	Employee
Edmondson, 1999	No dimension	Employee
Zani et al., 2001	Cognitive (perceived seriousness of problems in living environment), emotional (worried degree of negative events) and behavioral (behaviors to face feeling of unsafety)	Adolescent
Cong and An, 2004	Interpersonal security and certainty in control	–
Tynan, 2005	Self-psychological security and other psychological security	Employee
Dzhamalova et al., 2016	Senses and feelings, perception and evaluation of reality according to the criterion of dangerous-safe, and analysis and forecasting for a secure future	–
Yin, 1980	Risk estimation and severity evaluation of injury	Resident
Van der Wurff et al., 1989	Attractivity for crime, evil intent and power (feelings of self-assurance, control and confidence in meeting crime) and criminalizable space	Resident
Hale, 1996	Street crime, emotional security, physical security and property security	Resident
Vail, 1999	Property security, personal security, traffic security, medical security, food security and labor security	Resident
Rader, 2004	Emotive component (fear of crime), cognitive component (perceived risk), and behavioral component (constrained behaviors)	Resident
Xia and Wei, 2011	Economic security, interpersonal security, social security, environmental security, and survival security	Resident

Measurement

Some representative results of psychological security dimension and scale research are shown in **Table 1**. At present, there are few researchers paying attention to the measurement of the psychological security of urban residents. Most research is focused on measuring psychological safety in the workplace, in which individual-level studies of employees are mostly assessed using the Dyadic Psychological Safety Items designed by Tynan (2005). This scale includes two dimensions of self-psychological safety and other-psychological safety, with a total of 12 items. Team-level studies are mostly conducted using the Team Psychological Safety Scale (Edmondson, 1999). The scale contains seven self-evaluation items, and there are no separate dimensions. Most researchers have used a revised version of this scale (Pearsall and Ellis, 2011; Leroy et al., 2012; Hood et al., 2016). The Psychological Climate Scale developed by Brown and Leigh (1996) is widely used in organizational-level studies (Ogilvie et al., 2017; Ho et al., 2018) and includes measurement of supportive management, role clarity, contribution, recognition, self-expression, and challenges, consisting of 21 items. However, the dimension setting is applicable to only occupational sites but not to urban residents. Zani et al.'s, (2001) research on adolescents had similar problems.

Maslow (1942) developed the Psychological Security-Insecurity Questionnaire and believed that psychological security can be divided into three dimensions: safety; belongingness; and receiving love and affection. The Security Questionnaire developed by Cong and An (2004) includes two dimensions: interpersonal security and certainty in control. Both measurement tools and dimensions are widely used, but because the research subjects are not limited, the questionnaire must be adapted to specific situations. In recent years, some researchers have incorporated the perception of social reality into the structure of psychological security. For example, on the basis of the external perception of stable personality, Dzhamalova et al. (2016) believe that psychological security consists of senses and feelings, perception and evaluation of reality according to the dangerous-safe criterion, and analysis and forecasting for a secure future. The psychological security state at the city level is based on the individual psychological state and is influenced by environmental factors. Previous studies have mainly focused on the fear of crime (Yin, 1980; Van der Wurff et al., 1989; Rader, 2004). Hale (1996) combined emotional and social factors to divide psychological security into four dimensions: street crime, emotional security, physical security, and property security. Vail (1999) considered more social factors and constructed a six-dimensional structure including property security, personal security, traffic security, medical security, food security, and labor security. The Resident's Sense of Security Scale developed by Xia and Wei (2011) includes factors of economic security, interpersonal security, social security, environmental security, and survival security. This research incorporates elements from psychological factors, social factors, and environmental factors, but it is not comprehensive.

The previous scales measuring psychological security mostly focus on the multi-level security of the employees in the workplace. Other than this, the research focus of other scales has been diverse but scattered, and the degree of recognition is generally not high, and application field and scene are limited. A specific questionnaire to measure urban residents' psychological security is lacking. Therefore, it is important to develop a scale of urban residents' psychological security based on three-dimensional structure of psychology, society, and environment, which reflects social reality.

Thus, the connotation of psychological security of urban residents has changed over time, and the existing literature is lacking in terms of reflecting the comprehensive indicators of psychological, social, and environmental aspects. The development of the URPS scale has expanded the work in this field to some extent. Moreover, the grounded theory emphasizes the utilization of original data and fills the gap between theory and reality through methods such as literature review, interviewing, and coding, which can effectively address the defects in previous research in this field (Glaser et al., 1968). Consequently, we used a combination of qualitative and quantitative methods to develop the URPS scale, based on extensive literature research. We used the grounded theory to develop the initial scale and used the data collected through investigation questionnaires to quantitatively analyze the structure of the URPS scale.

INITIAL SCALE CONSTRUCTION BASED ON GROUNDED THEORY

Participants and Design

In order to extract the items for the initial UPRS scale, we conceptualized urban residents' psychological security and presented the specific performances of its structure. We obtained the original items using the following methods: (1) we conducted targeted interviews of urban residents and used recording software to reorganize, edit, and export the interviews. (2) We reviewed the existing literature and systematically analyzed the theory and empirical research results regarding security and psychological security to provide theoretical support for the scale.

The interviews did not include pre-set patterns or pre-assumptions but did consist of a specific outline. The outline was an auxiliary tool for us to guide interviewees by reviewing and describing relevant question, which is provided in **Table 2** below.

The questions listed in **Table 2** are only for reference. The interview was adjusted according to each specific situation. In addition to obtaining basic information, we conducted extended interviews depending on the interviewee reactions or answers.

Ethics Statement

This study was carried out in accordance with the principles of the Basel Declaration and recommendations of Ethical Codes of Consulting and Clinical Psychology of Chinese Psychological Society, Chinese Psychological Society. The protocol was approved by the Ethics Committee at the Department of Organizational and Behavioral Sciences, China University of Mining and Technology. All subjects gave written informed consent in accordance with the Declaration of Helsinki. Before the interview, the interviewees were told that they would be recorded and that we would fully respect their wishes.

Procedure

Based on the grounded theory and research requirements, we needed urban residents as the research subjects, with different educational backgrounds, different income levels, and mainly young and middle-aged people. Therefore, 25 interviewees were randomly selected through online recruitment. We conducted descriptive statistical analysis on the basic information of the interviewees. The results indicated that 52% were males, and 48% were females; 36% were between ages 22 and 30, 32% were between 31 and 40, and 32% were over 40; and 68% had received undergraduate education or above. In addition, our study included urban residents with different income levels and city. The sample is representative.

We converted the interview recordings into text, and on completion had obtained interview records of about 30,000 words. Eight respondents were randomly selected for theoretical saturation test, but their answers did not bring new information to the research, that is, the content was saturated in theory. The researchers read the original text content of the interview word by word, collected phrases about psychological security, and

TABLE 2 | Outline of interview on psychological security of urban residents.

Theme	Main content
Basic information	Gender, age, educational background, monthly income level, work place, nature of organization
The status of urban residents' psychological security	a. What do you think of the city you live in now? How does it feel to live in this city? b. What advantages do you think this city has? What are the shortcomings? c. Do you sometimes feel worried, anxious, panic or afraid in your daily life in this city? Can you give me an example?
The structure of urban residents' psychological security	a. Can you describe the situation and feelings when you feel safe in your daily life? b. Can you describe the situation and feelings when you feel safe at work?

TABLE 3 | Classification of the semantically similar items.

Original statements in the interview text	Conceptualization	Frequency
The environment is not very good, food is not safe, the network is not safe, and it is easy to be scammed by the Internet, all of these make me feel unsafe that living in this city; In foreign countries, people can carry guns and dare not go out in the middle of the night. I think China has done a very good job in this regard and I hope to increase the sanctions on robbery and theft; I am afraid to eat gutter oil, food is not safe, there are many pesticides and fertilizers	Social risk perception	25
I work too long every day and worry about my health; My job is very stable, my income is very guaranteed, and these are my sense of security; Large work pressure, fast life pace, I am busy every day. If I am not at work, I am on the way to work.	Occupational security	20
I am not an egoist, and I will get happiness by helping others; There is not much intimacy between people, everyone is very busy, and they don't want to take care of you; There are a few friends in my daily life. When I have a job, I can have a dinner party or go out to play with my colleagues	Interpersonal security	15
After graduating, I moved around several places and eventually returned to my hometown. To be honest, the economy here is not particularly developed, and the job opportunities are relatively small, but ultimately it is my hometown; The pace of life in this city is not fast, prices are not low, and the environment is not good. The reason why I stay here is that my family is here; This city is my home. No matter how long I play outside, I still have to come back	Urban belongingness	15
The air quality is not good, and the haze will pollute various things and damage the health; The haze is frequent, and the environment is relatively poor. Now, the city managers are trying to improve the situation, I hope that we can persevere to manage the environment, which is a benefit for the local people and the outsiders; Concerned about air quality, pollution problems in heavy industrial cities are very serious	Environmental pollution risk perception	14
Surrounding people live according to the established model, without any incitement or turmoil; Being able to have a safe environment to ensure that I can implement my plan and I will not interrupt my plan or fail to complete it due to some unexpected circumstances; I am worried that housing prices will become higher, I am anxious about the growth rate of wages, and I am afraid of the development of IT industry in Xuzhou. I am scared that 1 day I need to leave my hometown for a better life	Certainty in control	13
......	...	...

extracted conceptual labels from them. In order to ensure the objectivity of the label, the extracted statements were the original words of the interviewee.

After preliminary classification, 22 items were obtained from a total of 133 original statements. The researchers discussed the statements several times and decided to reclassify them according to their semantic similarity and delete ambiguous items, meaning that 126 statements remained. Due to the complexity of the 126 statements, the researchers combined and simplified them based on the literature review to form conceptual indicators. The specific classification is shown in **Table 3**.

An individual's accurate perception of self and future is crucial to their mental health (Taylor and Brown, 1988). Studies have indicated that interpersonal security can effectively promote the connection between the individual and the outside world, narrowing the boundaries between the inside and the outside of the group (Zhang et al., 2015). Simultaneously, when an individual lacks a sense of control over the future, anxiety, stress, and depression accompany this. Based on the item collection and primary research, "have like-minded friends" refers to "interpersonal security" and "have a safe environment to ensure the implementation of my plan" refers to "certainty in control." We summarized these statements about the safety perception of urban residents in relation to their own psychological status as "self-psychological security," resulting in the development of 6 scale items.

The degree of urban residents' sense of recognition with the city can directly reflect the urban integration degree of the inflow population. Having a good occupation is an important economic foundation and a spiritual pillar for urban people to live in the city and it is an important way to engage in social interaction and realize the individual's value. Research has indicated that social security (Foster et al., 2016; Prieto Curiel and Bishop, 2017), food security (Martin et al., 2016; Tseng et al., 2017), and related factors all have an impact on psychological security. We noted that there were a lot of emotional expressions about cities, society, and occupations in the collected statements. We classified them in detail. For example, "this city is my home" and "no matter how fun outside, you still have to come back" refer to "urban belongingness"; "I like my job very much" and "I am worried that I can't shoulder the pressure of work" refer to "occupational security"; "Afraid of being scammed by the internet" and "may not be well cared when I am old" refer to "social risk perception." Interestingly, we have found that technological advancement has increased people's negative psychological pressure while improving their quality of life. These statements refer to the aforementioned phenomenon in the following way: "technology is progressing too fast, it is difficult to adapt" and "it takes a long time to play games every day, sometimes I feel empty." refer to "technology risk perception." In consideration of our literature review, residents' sense of attachment to the living city, occupational stability, social risk perception, and negative perception of technology were summarized as "social environmental security," resulting in the development of 13 scale items.

In recent years, more and more people are paying increasing attention to the impact of environmental pollution and climate change on their health. Burnett et al. (2018), Chen et al. (2018), and Obradovich et al. (2018) have also found that air pollution and climate change not only threaten people's lives but also have significant positive impact on mental illness. Some of the statements show the public's close attention to the state of the environment. For example, "there is often smog, the environment is poor" and "the air is not good, the pollution problem is very serious." Consequently, we summarized "environmental pollution risk perception," "climate change risk perception," and "natural disaster risk perception" as "natural environmental security," which reflects urban residents' perceptions of their own surrounding environment, resulting in the development of five scale items.

Based on our review of prior studies and numerous discussions, several experts reorganized, classified, and extracted the expressions to develop a URPS scale that consisted of 24 items. The specific structure is shown in **Figure 2**. The purpose of this research was to enhance the theoretical logic and content validity of the assessment of urban residents' psychological security through qualitative research methods. In the next stage, quantitative research methods were used to present and examine the measure through obtaining empirical data.

QUANTITATIVE METHOD

Preliminary Survey and Extraction of the URPS Scale

Participants

The purpose of this preliminary survey was to evaluate the quality of the initial questionnaire, to purify and correct the items in the initial questionnaire, and to develop the formal URPS scale. In June 2018, we conducted preliminary surveys of residents in different urban areas. Firstly, through haphazard sampling, the research team members publicized and spread the network links of the online questionnaire on social platforms, and expanded the number and scope of the respondents by constantly forwarding links. Secondly, in order to make the distribution of the surveyed population in the demographic characteristics reasonable, stratified random sampling was adopted to distribute some questionnaires with the help of China's professional questionnaire survey website. Finally, we compared the selected demographics with the national demographics. Survey sample demographics conformed well to the national demographics. Meanwhile, to ensure resident's active participation, we provided cash rewards after completing the questionnaire.

409 questionnaires were collected. We deleted questionnaires with missing options or more than eight consecutive questions selecting the same option, and identified 304 valid questionnaires (74.3%). We conducted a descriptive statistical analysis of the preliminary survey samples and found that: 47.7% were males and 52.3% were females; the distribution of age was the reflection of the distribution in social reality, with 24.3% of the individuals below the age of 25, 35.2% between 26 and 35, 28.6% between 36 and 45, and 11.8% older than 45. The samples were suitably representative.

Procedure

First, we performed a reliability test on the initial scale. (1) Cronbach's α coefficient was used to judge the overall credibility of the scale. After reverse scoring of the items 1–6, 13–15, and 17–19, the results showed that the Cronbach's α value of the URPS scale was 0.788, indicating that the overall reliability of the scale was acceptable. (2) Project analysis was used to determine the credibility of every item, including a total of four methods: (1) Descriptive statistical analysis. The descriptive statistical data for each item was used to assess the basic quality of that item, and there were no low-discrimination items with standard deviations less than 0.75. (2) Extreme group test. Among the 304 residents surveyed, we selected 27% of the highest total scores and 27% of lowest total scores, that is, a total of 167 people whose score was higher than 82 points or below 167 points as extreme groups, and we performed independent sample t-tests for the extreme groups. The t-test values all reached a significance level of 0.05, indicating that all the items can effectively identify the high and low scores. (3) Correlation test. Among the 24 questions in the scale, all the items were significantly correlated with the total score of the scale. (4) Cronbach's α value test. The data showed that the overall credibility value of the scale would decrease after deleting any item. Thence, after the project analysis, there were still 24 items in the URPS scale.

Second, we conducted principal component analysis on the 24 items. During the analysis, we removed any item with a factor load value less than 0.5 or a cross load value over 0.4. After multiple factor analysis, the 7th, 16th, 17th, and 18th items were deleted, and a well-discriminating factor structure was obtained. Consequently, we developed a URPS scale with 20 items.

Finally, based on the feedback from some interviewees and the re-discussion of experts, we improved the linguistic expression of the scale items, thereby further improving the accuracy and clarity of the scale expression and improving the content validity of the scale.

In summary, we improved the quality of the initial scale through conducting a pre-study assessment and a formal survey using the URPS scale consisting of 20 items (see **Supplementary Appendix**). The scale was used in the formal survey.

Formal Survey and Structural Analysis of the URPS Scale

Data Collection

In February 2019, we collected data using questionnaires. A total of 1,036 formal questionnaires was sent out and 985 copies were returned, of which 802 were valid, and the effective recovery rate was 77.4%. The specific distribution of the sample is shown in **Table 4**.

Exploratory Factor Analysis

Exploratory factor analysis was performed on the optimized scale using SPSS 19.0 with half of the data (N = 401). As the KMO value of the scale was 0.803 > 0.8, the Bartlett test was passed (p = 0.000 < 0.001), indicating that the variables correlated

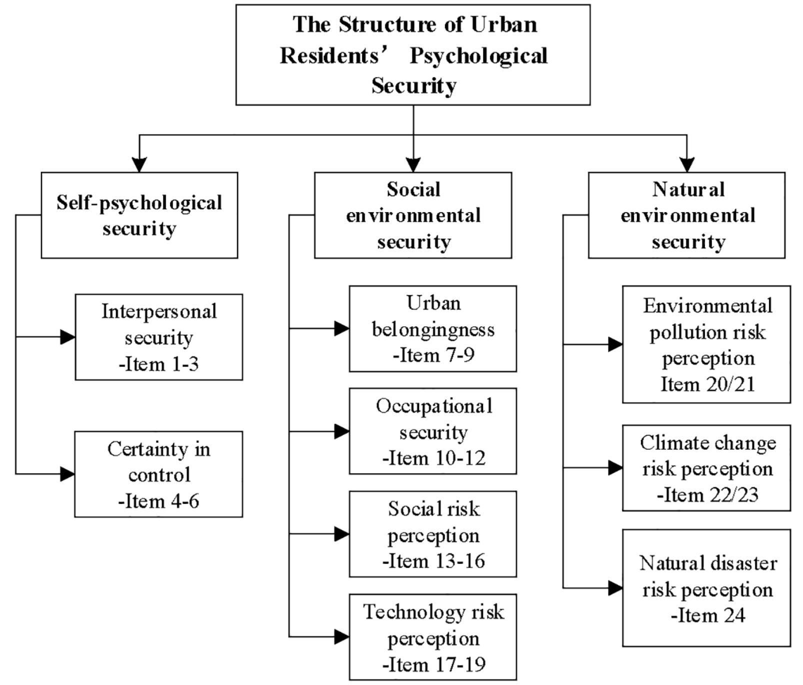

The Structure of Urban Residents' Psychological Security
Self-psychological security
Social environmental security
Natural environmental security
Interpersonal security -Item 1-3
Certainty in control -Item 4-6
Urban belongingness -Item 7-9
Occupational security -Item 10-12
Social risk perception -Item 13-16
Technology risk perception -Item 17-19
Environmental pollution risk perception Item 20/21
Climate change risk perception -Item 22/23
Natural disaster risk perception -Item 24

FIGURE 2 | The structure of urban residents' psychological security.

and were suitable for factor analysis. The principal component analysis method and varimax orthogonal rotation were used to obtain the factor load matrix as shown in **Table 5**. According to the Kaiser criterion, we extracted four factors with eigenvalues higher than 1, and the accumulated variance explanation rate of these four factors was 52.5%.

Combining the items of each dimension and the analysis of the related literature, we named and defined the four scale factors explored by principal component analysis as follows:

(1) "Natural environmental security" (5 items), is the overall perception of urban personnel on the natural environment state of living cities;
(2) "Self-psychological security" (6 items), is the safety expectation of urban personnel for future life and interpersonal relationships according to their past life experience;
(3) "Social security" (5 items), as the individual's sense of stability and belonging within urban life.
(4) "Social environmental risk perception" (5 items), is the individual's overall perception of social risk in urban life.

Confirmatory Factor Analysis

We used the other half of the data sample (N = 401) to test how well the conceptual model obtained by the exploratory factor analysis fit the actual observed data. In order to better verify the accuracy of the model, four competition models are proposed below, which are compared with the results of the above exploratory factor analysis.

We set Four alternative models:

M1: single factor model in which we hypothesized that the 20 items had a common latent variable: URPS.
M2: two-factor model in which we hypothesized that 11 items from natural environmental security and self-psychological security would have common latent variables, and 9 items from social environmental security have common latent variables.
M3: the three-factor model in which we hypothesized that 5 items of natural environmental security would have common latent variables, 6 items of self-psychological security would have common latent variables, and 9 items of "social security" and "social environmental risk perception" would have common latent variables: social environmental security.
M4: four-factor model in which according to the results of exploratory factor analysis, we hypothesized that the four factors of "social security" and "social risk perception" in natural environment security, self-psychological security, and social environmental security would be factors in this model.

For each of the above models, we used each factor as the latent variable and the corresponding items as the observational variables to perform confirmatory factor analysis, and the model fit results are shown in **Table 6**. The fit results for M1, M2, and M3 were not ideal. The GFI, AGFI, NFI, CFI, TLI, and IFI for three models were all less than 0.9, and the RMSEA value of M1 and M2 were both greater than

TABLE 4 | Sample distribution.

Sex	*N*	Age	*N*	Marital status	*N*
Male	389	<18	1	Married	497
Female	413	18–25	233	Spinsterhood	292
Nature of organization	***N***	26–35	252	Else	13
Government	22	36–45	215	**Monthly income (RMB)**	***N***
Public institution	152	46–55	85	<3000	108
State-owned company	151	>55	16	3000–5000	169
Collectively ownership institution	12	**Diploma level**	***N***	5001–8000	225
Private company	255	Senior high school and following	40	8001–10000	128
Joint venture company	53	Junior college	96	10001–20000	131
Sino-foreign joint company	12	Bachelor's degree	561	20001–50000	38
Foreign-funded company	39	Master's degree	85	>50000	3
Joint-stock company	30	Ph.D. and above	20		
Else	73				

0.1. The χ^2/df of the M4 model was 2.009, which is the smallest when compared to the other three models, and the GFI, AGFI, CFI, TLI, and IFI of M4 were all greater than 0.9. Therefore, we considered that the M4 model was the optimal first-order model.

However, there were still some indicators that did not meet expectations. We revised the model parameters and released the variance coefficients with a correction index greater than 10, as shown in **Table 7**.

After twice model corrections, the GFI, AGIF, NFI, TLI, and CFI values were all greater than 0.9, the RMSEA value was below 0.05, and the χ^2/df value was 1.601, indicating that the data fit well with the model, and all indicators achieved good results. Thus, the URPS model had an ideal fit. The standardized path diagram is provided in **Figure 3**.

Reliability and Validity

The evaluation of the reliability of the scale mainly included two levels of the overall credibility of the scale and the credibility of the latent variables. The Cronbach's α value (>0.7) was used to test the overall credibility of the scale and the credibility of the latent variable was tested by both the Cronbach's α value and CR value. The analysis showed that the overall Cronbach's α value of the URPS scale was 0.773, indicating that the overall credibility of the scale is reliable. The CR value of each latent variable was between 0.75 and 0.9, and the Cronbach's α values for each latent variable were 0.828, 0.806, 0.686, and 0.670, respectively. Since each principal component is not measured as a single variable and has fewer items, the reliability values were within acceptable limits and the scale passed the reliability test.

The evaluation of the validity of the scale mainly included two aspects: content validity and structural validity. The content validity was ascertained using qualitative methods. The verification of structural validity examines the convergence validity and discriminant validity of the scale. We strictly followed standard scale development procedures. We conducted a large scale literature review, collected initial items through in-depth interviews based on grounded theory, invited management experts to discuss the design of the questionnaire repeatedly, and a pre-study utilizing 304 questionnaires, so the content validity of this scale is reliable. In addition, the standardized load of 20 scale items at the corresponding latent variables was greater than 0.5 and reached the level of statistical significance, and the corresponding AVE value was between 0.45 and 0.65, which satisfies AVE > 045, indicating good convergence validity of the scale. The square root of the AVE of the latent variable was greater than the correlation coefficient between the latent variables, indicating that the potential structural discrimination of the variable was better. The scale passed the validity test. The specific analysis is shown in **Table 8**.

Criterion Correlation Validity

We used the psychological security of urban residents measured by single global rating as the criterion. Respondents answered one question about their general feeling of security in urban life: "Based on your daily life in the city, what do you think your psychological security score is?" The question was scored on a Likert scale, in which 1 means "very unsafe," and 5 means "very safe."

Harman single factor test was carried out on 21 items including the URPS scale and the item of single global rating. The results showed that 21 items were automatically divided into 4 factors instead of one factor, and the variance contribution rate of the first main factor was 19.706%, which was much less than 40%. It can be seen that the common method bias has no significant interference with the criterion correlation validity test.

As shown in **Table 9**, there was a significant positive correlation between the mean value of the URPS scale and the results measured by single global rating. The four main factors scores of the scale were also significantly correlated with the score of psychological security, with a correlation coefficient between 0.2 and 0.4. To further investigate the explanatory power of the scale regarding psychological security, we conducted regression analysis. First, gender, age, education background and income as demographic variables were used as variables in model 1, and the adjusted R^2 was only 0.023, thus indicating that demographic variables explained only 2.3% of psychological security. Then, four main factors were included in model 2, and the adjusted R^2 was 0.219, and the F value was significant at the 0.001 level, thus indicating that the four factors of the scale had a significant positive prediction effect on psychological security. Finally, the mean value of the scale was included in model 3, and the adjusted R^2 was 0.191, and the F value was significant at 0.001, thus indicating that the mean value of the scale was able to significantly positively predict the psychological security of urban residents. Therefore,

TABLE 5 | Exploratory factor analysis results.

Item	Commonality	Factor			
		S1	S2	S3	S4
CCRP-1	0.688	0.819			
EPRS-2	0.646	0.793			
CCRP-2	0.604	0.771			
NDRP	0.566	0.728			
EPRS-1	0.474	0.678			
IS-1	0.592		0.763		
IS-3	0.590		0.739		
IS-2	0.570		0.736		
CC-1	0.520		0.664		
CC-2	0.465		0.619		
CC-3	0.460		0.558		
SRP-1	0.568			0.744	
SRP-2	0.587			0.720	
SRP-3	0.469			0.653	
TRP	0.418			0.571	
UB-1	0.480				0.691
OS-1	0.530				0.683
OS-3	0.464				0.633
OS-2	0.405				0.608
UB-2	0.403				0.596
Factor name		**Natural environmental security**	**Self-psychological security**	**Social environmental risk perception**	**Social security**
Eigenvalues		4.054	3.268	1.881	1.297
Factor variance contribution%		20.269	16.339	9.404	6.486
Accumulated variance contribution%		20.269	36.607	46.012	52.498

CCRP, climate change risk perception; EPRS, environmental pollution risk perception; NDRP, natural disaster risk perception; IS, interpersonal security; CC, certainty in control; SRP, social risk perception; TRP, technology risk perception; UB, urban belongingness; OS, occupational security.

TABLE 6 | Major fitting degree indices of urban residents' psychological security.

Model	χ^2	*df*	χ^2/df	GFI	AGFI	NFI	CFI	TLI	IFI	RMSEA
M1: Single factor model	1525.020	170	8.971	0.618	0.529	0.342	0.363	0.288	0.369	0.141
M2: Two-factor model	1235.130	169	7.308	0.685	0.609	0.467	0.499	0.436	0.503	0.126
M3: Three-factor model	543.978	167	3.257	0.866	0.831	0.765	0.823	0.798	0.825	0.075
M4: Four-factor model	329.439	164	2.009	0.923	0.901	0.858	0.922	0.910	0.923	0.050

the URPS scale developed in this paper had good criterion correlation validity.

DISCUSSION AND CONCLUSION

Discussion

We attempted to integrate the conceptual connotation of URPS by borrowing the elements from diverse literature. A scale comprising three dimensions (psychology, society and environment) was developed. The measurement of URPS from the dimension of self-psychological security, natural environmental security and social environmental security has objective rationality, thus authentically and explicitly demonstrating the current state of URPS. For example, Zhang (2007) have divided the feeling of security of residents into psychological security, social security, economic security government security and environmental security. However, Zhang did not consider the influence of climate change risk perception, technology risk perception, urban belongingness and other factors. Moreover, although the survey was conducted in China, the scale is not only suitable for developing countries that have achieved rapid economic growth at the expense of the environment, such as China and India, but also is suitable for developed countries that have strict environmental requirements, such the European Union and the United States.

The dimension of self-psychological security was established on the basis of previous studies, including interpersonal security

TABLE 7 | Overall fitting degree indices of each modification.

		Initial model fitting	Release e16-e17	Release e10-e11	Assessment
Absolute fitting index	X^2	329.439, *df* = 164 *P* = 0.000	282.753, *df* = 163 *P* = 0.000	259.410, *df* = 162 *P* = 0.000	Great
	GFI	0.923	0.933	0.938	Great
	RMR	0.065	0.062	0.062	Good
	RMSEA	0.050	0.043	0.039	Great
Relative fitting index	AGFI	0.901	0.914	0.920	Great
	NFI	0.858	0.878	0.888	Good
	TLI	0.910	0.934	0.946	Great
	CFI	0.922	0.944	0.954	Great

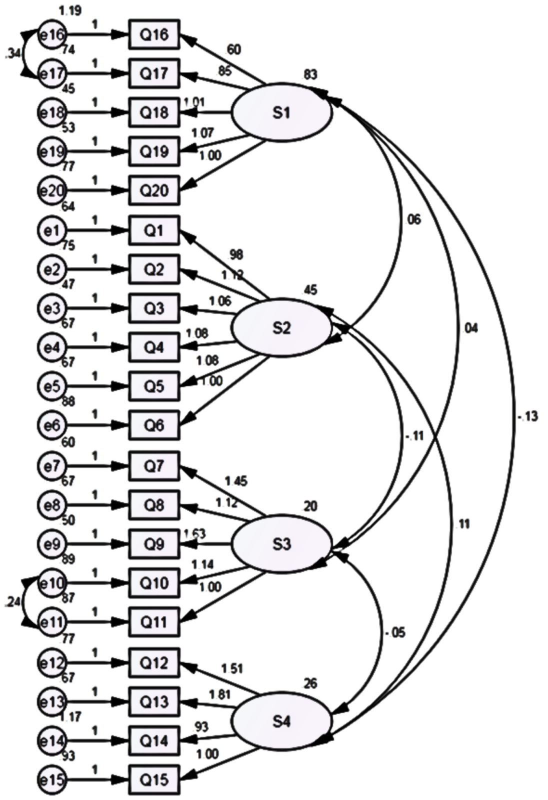


FIGURE 3 | Estimations of the standardized path coefficient of the final confirmatory factor model.

and certainty in control. Gunn et al. (2014) have found that interpersonal distress decreases people's sense of security, in agreement with the results of this paper. People who cannot trust others and who avoid others as much as possible in interpersonal communication cannot accept themselves well and tend to make negative comments about themselves, thereby affecting their psychological security (Cong and An, 2004). Steptoe et al. (2007) and Chou and Chi (2001) believe that a low sense of control is associated with depressive symptoms, thus supporting the factor of certainty in control in this paper. People with a lower sense of control often feel that their lives are out of control or a mess, or that they cannot cope with life's unexpected problems; consequently, they are always in a state of insecurity. Therefore, we believe that interpersonal security and certainty in control can effectively reflect the state of URPS.

The dimension of natural environmental security includes air pollution risk perception, climate change risk perception and natural disaster risk perception. To date, pollution and climate change in environmental factors have rarely been considered in the development of the psychological security scale of urban residents; this consideration can be regarded as an innovation of this paper. Jacquemin et al. (2007), Lucchini et al. (2012), and Sucker et al. (2008) have found that exposure to pollution stimulates nerves in the brain, thus causing negative emotions such as worry, anxiety, tension and aggression. Having negative emotions for a long time increase individuals' sense of dissatisfaction and vigilance, and affects their sense of security. Although there is still controversy in the public opinion on whether global climate change exists and whether it can threaten human life (Leiserowitz, 2005; Weber and Stern, 2011), the risk perception of extreme cold and hot weather, sea level rise and food loss brought by climate change, are real threats to people's psychological security. If an individual has experienced natural disasters such as tsunamis, earthquakes, floods or tornadoes, a trauma will result that is difficult to heal for individual psychology (Weinstein et al., 2000; Williams, 2006). People who have experienced trauma show severe stress reactions over a long period. They are extremely sensitive to external threats and may have long-term mental disorders that severely affect their psychological security. Therefore, urban residents' perception of the risks of air pollution, climate change and natural disasters play a key role in the URPS.

The dimension of social environmental security includes two factors: social security and social risk perception. Social security includes urban belongingness and occupational security. Social risk perception includes medical, pension, food and technology risk perception. The factor of urban belongingness is the reflection of psychological security in the urban context; consideration of this factor is another unique feature of this paper, as compared with the general psychological security scale. The sense of city identity increases residents' living satisfaction and brings about positive psychological expectations (Zenker and Petersen, 2014). The economic factor is the guarantee of individual security, and the main economic security of urban residents is based on having a stable occupation. Whether the city is able to provide satisfactory jobs is a key issue for urban residents (Vieitez et al., 2001), and also are the main factors in this paper. Moreover, Bodie et al. (2009), Hesketh et al. (2012), Yan (2012), Gille et al. (2015), and Wu et al. (2017) believe that medical supervision, pension resources, food safety and other issues have caused urban residents to have negative emotions, such as anxiety. Therefore, urban belongingness, occupational status and social factors can directly influence the psychological security of urban residents.

In addition, we also found that the negative spillover effects brought about by the development of technology affect the individual's mental health. This can be considered as a new development in the field of psychological security structures

TABLE 8 | Reliability and validity test of latent variables.

	Natural environmental security	Self-psychological security	Social environmental risk perception	Social security
Natural environmental security	0.796*			
Self-psychological security	0.391	0.740*		
Social environmental risk perception	−0.066	0.330	0.679*	
Social security	0.371	−0.458	−0.052	0.711*
Cronbach's α	0.828	0.806	0.686	0.670
CR	0.8733	0.8493	0.7914	0.7762
AVE	0.6343	0.5472	0.4606	0.5060

**indicates the square root of AVE value.*

TABLE 9 | Correlation coefficient and regression results.

Variable	Natural environmental security	Self-psychological security	Social environmental risk perception	Social security	Mean of URPS
Psychological security	0.363***	0.213***	0.261***	0.363***	0.427***

	Psychological security		
	Model 1	**Model 2**	**Model 3**
Constant	3.382***	1.490***	1.699***
Gender	−0.108*	−0.117**	−0.112*
Age	−0.011	−0.016	−0.014
Education	0.085*	0.08*	0.080*
Income	0.045*	−0.016	0.001
Natural environmental security		0.174***	
Self-psychological security		0.117***	
Social environmental risk perception		0.311***	
Social security		0.051*	
Mean of URPS			0.594***
F	5.801***	29.042***	38.859***
R^2	0.028	0.227	0.196
ΔR^2	0.023	0.219	0.191

*$p<0.05$; **$p<0.01$; ***$p<0.001$.

of urban residents. Most previous research has focused on the benefits of technological advances, such as general increases in productivity and quality of life. Internet technology is widely used worldwide and can connect people across distances and enhance interpersonal communication, such as cross-border communication. However, we found in the interviews that the rapid updating of technology makes elderly people or those with low adaptability fear being abandoned by the times, and their unfamiliarity with the Internet leads to their fear of being swindled and robbed. Young people are more familiar with the online environment, but they spend too much time communicating on the Internet and thus neglect the real world. Moody (2001) as found that the massive use of Internet technology has caused some people to be lonely and socially isolated in the real world, in agreement with our findings from this study. Some researchers believe that lonely individuals use the Internet more to modulate negative moods and obtain emotional support (Morahan-Martin and Schumacher, 2003). This paper argues that individuals too immersed in the semi-virtual world of the Internet will expend a large amount of emotional energy, leading to emotional exhaustion and interpersonal alienation in the real world. Excessive feelings of loneliness and alienation reduce the individual's psychological security.

Conclusion

(1) We first conducted in-depth interviews with 25 urban residents, and combined with a literature review, developed the initial URPS scale consisting of 24 items through qualitative analysis. Subsequently, we used project analysis and principal component analysis to purify the scale and verify the structure of scale, using 304 pre-survey questionnaires, and then developed a formal survey URPS scale, with 20 items.

(2) A total of 802 formal questionnaires were collected. Through principal component analysis of 401 samples, "natural environmental security," "self-psychological security" and "social environmental security" (including

social security and social risk perception) were obtained. The KMO value was 0.803, which is greater than 0.7, the significance was 0.000, and the cumulative variance of the four factors was 52.498%. We performed confirmatory factor analysis on the other half of the data and found that the M4 model was superior to the other three models. Simultaneously, because some indicators were not excellent, the model parameters were corrected. The GFI, AGIF, TLI and CFI values of the modified model were 0.938, 0.920, 0.946, 0.954, respectively. The RMSEA value was 0.039, and the χ^2/df value was 1.601. In summary, the good range showed that the URPS model had an ideal fit.

(3) Reliability test and validity test were performed on the developed scale. Cronbach's α value of the overall credibility of the scale was 0.773, which is higher than 0.7, and Cronbach's α values for each latent variable were 0.828, 0.806, 0.686, and 0.670, respectively. The CR values for each latent variable were 0.8733, 0.8493, 0.7914, and 0.7762, separately. On the basis of accepted standards, the scale passed the reliability test. The scale was developed in strict accordance with recommended procedures and the development was scientific and rigorous. Analyses demonstrated that the content validity was reliable. The standardized loads of 20 scale items at the corresponding latent variables were all greater than 0.5, and the corresponding AVE values were 0.6343, 0.5472, 0.4606, and 0.5060, respectively, all of which were above 0.45. The scale convergence validity was high, and the square root of the AVE of the latent variable was greater than the correlation coefficient between the latent variables. In addition, the degree of potential variable structural discrimination was better. Importantly, the scale also passed the validity test. In the criterion correlation validity test, the correlation coefficient between the results of psychological security of urban residents measured by single global rating and the mean of the URPS scale is 0.427, and the correlation coefficients between it and the mean of each dimension were 0.363, 0.213, 0.261, and 0.363, respectively. Regression analysis showed that URPS scale was able to significantly predict psychological security at the 0.001 level, with ideal criterion validity.

LIMITATIONS AND FUTURE STUDIES

There are some limitations in this study: (1) there are regional limitations in the choice of samples. Although the samples used were representative of most demographic variables when taking into account the economically developed and underdeveloped regions of China but there are still some areas that were not involved in this study, and there is no distinction between scales for different levels of urban development. (2) The focus of the research was on urban residents, so a large number of rural residents who complete the questionnaires were deleted, and this led to the lack of a comparative analysis between rural and urban residents. (3) The main contribution of this study was to develop a psychological security scale for urban residents, which has not been empirically tested. Therefore, it is necessary for the scale to be further verified, revised, and improved upon in future research.

Owing to the limitations of the development site, the validity of the scale was verified only in China. We expect to use this scale to measure and compare the psychological security of urban residents in different countries and cities in the future, and to verify that the URPS scale is applicable to different countries and regions. Next, we will conduct a large sample investigation by using the URPS scale. Then, we will analyze the differences in dimensions/variables among different regions and determine whether economic development, environmental pollution and technological development of different regions have significant differences in the four major factors, on the basis of the sample data. At the same time, urban residents' psychological security can be used as a mediator to study the resident turnover rate, sense of city integration and urban crime rate to improve city management level and city attraction.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

ETHICS STATEMENT

This study was carried out in accordance with the principles of the Basel Declaration and recommendations of Ethical Codes of Consulting and Clinical Psychology of Chinese Psychological Society, Chinese Psychological Society. The protocol was approved by the Ethics Committee at the Department of Organizational and Behavioral Sciences, China University of Mining and Technology. All subjects gave written informed consent in accordance with the Declaration of Helsinki. Before the interview, the interviewees were told that they would be recorded and that we would fully respect their wishes.

AUTHOR CONTRIBUTIONS

JW analyzed the data and wrote the manuscript. RL designed the framework of this manuscript. HC obtained the data and provided suggestions for improvement. QL made a major contribution to the manuscript revision process.

FUNDING

This study was financially supported by funds from the Key Project of National Social Science Foundation of China (No. 18AZD014) and the National Natural Science Foundation of China (No. 71874188).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02423/full#supplementary-material

REFERENCES

Astell-Burt, T., Feng, X., Kolt, G. S., and Jalaludin, B. (2015). Does rising crime lead to increasing distress? longitudinal analysis of a natural experiment with dynamic objective neighbourhood measures. *Soc. Sci. Med.* 138, 68–73. doi: 10.1016/j.socscimed.2015.05.014

Bodie, Z., Detemple, J., and Rindisbacher, M. (2009). Life-cycle finance and the design of pension plans. *Annu. Rev. Financ. Econ.* 1, 249–286. doi: 10.1146/annurev.financial.050708.144317

Brown, S. P., and Leigh, T. W. (1996). A new look at psychological climate and its relationship to job involvement, effort, and performance. *J. Appl. Psychol.* 81:358. doi: 10.1037/0021-9010.81.4.358

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., et al. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proc. Natl. Acad. Sci.* 115, 9592–9597. doi: 10.1073/pnas.1803222115

Carmeli, A., and Gittell, J. H. (2009). High-quality relationships, psychological safety, and learning from failures in work organizations. *J. Organ. Behav.* 30, 709–729. doi: 10.1002/job.565

Carter, K. N., Kruse, K., Blakely, T., and Collings, S. (2011). The association of food security with psychological distress in new zealand and any gender differences. *Soc. Sci. Med.* 72, 1463–1471. doi: 10.1016/j.socscimed.2011.03.009

Chen, S., Oliva, P., and Zhang, P. (2018). *Air Pollution and Mental Health: Evidence from China (No. w24686).* Report no. NBER Working Paper No. 246868, Cambridge, MA: National Bureau of Economic Research.

Chesley, N. (2005). Blurring boundaries? linking technology use, spillover, individual distress, and family satisfaction. *J. Mar. Fam.* 67, 1237–1248. doi: 10.1111/j.1741-3737.2005.00213.x

Chou, K. L., and Chi, I. (2001). Stressful life events and depressive symptoms: social support and sense of control as mediators or moderators? *Int. J. Aging Human Dev.* 52, 155–171. doi: 10.2190/9C97-LCA5-EWB7-XK2W

Cong, Z., and An, L. (2004). Developing of security questionnaire and its reliability and validity. *Chin. Men. Health J.* 18, 97–99. doi: 10.3321/j.issn:1000-6729.2004.02.010

Davila, J., and Sargent, E. (2003). The meaning of life (events) predicts changes in attachment security. *Personal. Soc. Psychol. Bull.* 29, 1383–1395. doi: 10.1177/0146167203256374

Demir, M. (2008). Sweetheart, you really make me happy: romantic relationship quality and personality as predictors of happiness among emerging adults. *J. Happiness Stud.* 9, 257–277. doi: 10.1007/s10902-007-9051-8

Doherty, T. J., and Clayton, S. (2011). The psychological impacts of global climate change. *Am. Psychol.* 66:265. doi: 10.1037/a0023141

Dzhamalova, B. B., Magomedov, G. B., Amirkhanov, A. A., Ramazanova, P. K., and Suleymanov, B. B. (2016). Anthropological mechanisms of self-management of personality behavior. *Int. Rev. Manag. Market.* 6, 383–389.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Admin. Sci. Q.* 44, 350–383. doi: 10.2307/2666999

Edmondson, A. C., and Lei, Z. (2014). Psychological safety: the history, renaissance, and future of an interpersonal construct. *Annu. Rev. Organ. Psychol. Organ. Behav.* 1, 23–43. doi: 10.1146/annurev-orgpsych-031413-091305

Evans, G. W. (2003). The built environment and mental health. *J. Urban Health* 80, 536–555. doi: 10.1016/B978-0-444-52272-6.00006-4

Foster, S., Hooper, P., Knuiman, M., and Giles-Corti, B. (2016). Does heightened fear of crime lead to poorer mental health in new suburbs, or vice versa? *Soc. Sci. Med.* 168, 30–34. doi: 10.1016/j.socscimed.2016.09.004

Gámez-Guadix, M., and Calvete, E. (2016). Assessing the relationship between mindful awareness and problematic internet use among adolescents. *Mindfulness* 7, 1281–1288. doi: 10.1007/s12671-016-0566-0

Garland, D. (2000). The culture of high crime societies. *Br. J. Criminol.* 40, 347–375. doi: 10.1093/bjc/40.3.347

Gille, F., Smith, S., and Mays, N. (2015). Why public trust in health care systems matters and deserves greater research attention. *J. Health Serv. Res. Policy* 20, 62–64. doi: 10.1177/1355819614543161

Glaser, B. G., Strauss, A. L., and Strutzel, E. (1968). The discovery of grounded theory; strategies for qualitative research. *Nurs. Res.* 17:364. doi: 10.1097/00006199-196807000-00014

Gunn, H. E., Troxel, W. M., Hall, M. H., and Buysse, D. J. (2014). Interpersonal distress is associated with sleep and arousal in insomnia and good sleepers. *J. Psychosom. Res.* 76, 242–248. doi: 10.1016/j.jpsychores.2013.11.010

Hale, C. (1996). Fear of crime: a review of the literature. *Int. Rev. Victimol.* 4, 79–150. doi: 10.1177/026975809600400201

Hart, J. (2014). Toward an integrative theory of psychological defense. *Pers. Psychol. Sci.* 9, 19–39. doi: 10.1177/1745691613506018

Hart, J., Shaver, P. R., and Goldenberg, J. L. (2005). Attachment, self-esteem, worldviews, and terror management: evidence for a tripartite security system. *J. Personal. Soc. Psychol.* 88:999. doi: 10.1037/0022-3514.88.6.999

Hesketh, T., Wu, D., Mao, L., and Ma, N. (2012). Violence against doctors in china. *BMJ* 345:e5730. doi: 10.1136/bmj.e5730

Ho, V. T., Kong, D. T., Lee, C. H., Dubreuil, P., and Forest, J. (2018). Promoting harmonious work passion among unmotivated employees: a two-nation investigation of the compensatory function of cooperative psychological climate. *J. Vocat. Behav.* 106, 112–125. doi: 10.1016/j.jvb.2018.01.005

Hood, A. C., Bachrach, D. G., Zivnuska, S., and Bendoly, E. (2016). Mediating effects of psychological safety in the relationship between team affectivity and transactive memory systems. *J. Organ. Behav.* 37, 416–435. doi: 10.1002/job.2050

Hu, Y., Zhu, L., Li, J., Maguire, P., Zhou, M., Sun, H., et al. (2018). Exploring the influence of ethical leadership on voice behavior: how leader-member exchange, psychological safety and psychological empowerment influence employees' willingness to speak out. *Front. Psychol.* 9:1718. doi: 10.3389/fpsyg.2018.01718

Inoue, A., Kawakami, N., Eguchi, H., and Tsutsumi, A. (2016). Buffering effect of workplace social capital on the association of job insecurity with psychological distress in Japanese employees: a cross-sectional study. *J. Occup. Health* 16, 460–469. doi: 10.1539/joh.16-0129-OA

Jacquemin, B., Sunyer, J., Forsberg, B., Götschi, T., Bayer-Oglesby, L., Ackermann-Liebrich, U., et al. (2007). Annoyance due to air pollution in europe. *Int. J. Epidemiol.* 36, 809–820. doi: 10.1093/ije/dym042

Jia, J., Li, D., Li, X., Zhou, Y., Wang, Y., and Sun, W. (2017). Psychological security and deviant peer affiliation as mediators between teacher-student relationship and adolescent internet addiction. *Comput. Hum. Behav.* 73, 345–352. doi: 10.1016/j.chb.2017.03.063

Kagan, J. (2009). Loneliness: human nature and the need for social connection. *Am J. Psychiatry* 166, 375–376. doi: 10.1176/appi.ajp.2008.08091320

Leiserowitz, A. A. (2005). American risk perceptions: is climate change dangerous? *Risk Anal.* 25, 1433–1442. doi: 10.1111/j.1540-6261.2005.00690.x

Leroy, H., Dierynck, B., Anseel, F., Simons, T., Halbesleben, J. R., McCaughey, D., et al. (2012). Behavioral integrity for safety, priority of safety, psychological safety, and patient safety: a team-level study. *J. Appl. Psychol.* 97:1273. doi: 10.1037/a0030076

Lu, H., and Long, R. (2018). Smog besieging the city: the influence mechanism of smog perception on talents flow tendency under dual channel perspective. *Econ. Manage. J.* 40, 104–124. doi: 10.19616/j.cnki.bmj.2018.11.007

Lucchini, R. G., Dorman, D. C., Elder, A., and Veronesi, B. (2012). Neurological impacts from inhalation of pollutants and the nose–brain connection. *Neurotoxicology* 33, 838–841. doi: 10.1016/j.neuro.2011.12.001

Martin, M. S., Maddocks, E., Chen, Y., Gilman, S. E., and Colman, I. (2016). Food insecurity and mental illness: disproportionate impacts in the context of perceived stress and social isolation. *Public Health* 132, 86–91. doi: 10.1016/j.puhe.2015.11.014

Maslow, A. H. (1942). The dynamics of psychological security-insecurity. *J. Personal.* 10, 331–344. doi: 10.1111/j.1467-6494.1942.tb01911.x

Mikulincer, M., and Shaver, P. R. (2007). Boosting attachment security to promote mental health, prosocial values, and inter-group tolerance. *Psychol. Inq.* 18, 139–156. doi: 10.1080/10478400701512646

Moody, E. J. (2001). Internet use and its relationship to loneliness. *CyberPsychol. Behav.* 4, 393–401. doi: 10.1089/109493101300210303

Morahan-Martin, J., and Schumacher, P. (2003). Loneliness and social uses of the Internet. *Comput. Hum. Behav.* 19, 659–671. doi: 10.1016/S0747-5632(03)00040-2

Obradovich, N., Migliorini, R., Paulus, M. P., and Rahwan, I. (2018). Empirical evidence of mental health risks posed by climate change. *Proc. Natl. Acad. Sci. U.S.A.* 115, 10953–10958. doi: 10.1073/pnas.1801528115

Ogilvie, J., Rapp, A., Bachrach, D. G., Mullins, R., and Harvey, J. (2017). Do sales and service compete? the impact of multiple psychological climates on frontline employee performance. *J. Pers. Sell. Sales Manag.* 37, 11–26. doi: 10.1080/08853134.2016.1276398

Oishi, S., and Kesebir, S. (2015). Income inequality explains why economic growth does not always translate to an increase in happiness. *Psychol. Sci.* 26, 1630–1638. doi: 10.1177/0956797615596713

Pearsall, M. J., and Ellis, A. P. (2011). Thick as thieves: the effects of ethical orientation and psychological safety on unethical team behavior. *J. Appl. Psychol.* 96:401. doi: 10.1037/a0021503

Prieto Curiel, R., and Bishop, S. (2017). Modelling the fear of crime. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 473:20170156. doi: 10.1098/rspa.2017.0156

Probst, T. M. (2002). "The impact of job insecurity on employee work attitudes, job adaptation, and organizational withdrawal behaviors," in *The Psychology of Work: Theoretically Based Empirical Research*, eds J. M. Brett, and F. Drasgow, (Mahwah, NJ: Lawrence Erlbaum Associates Publishers), 141–168.

Rader, N. E. (2004). The threat of victimization: a theoretical reconceptualization of fear of crime. *Sociol. Spectr.* 24, 689–704. doi: 10.1080/02732170490467936

Ross, C. E., and Hill, T. D. (2013). Reconceptualizing the association between food insufficiency and body weight: distinguishing hunger from economic hardship. *Sociol. Pers.* 56, 547–567. doi: 10.1525/sop.2013.56.4.547

Sekulova, F., and Van den Bergh, J. C. J. M. (2016). Floods and happiness: empirical evidence from bulgaria. *Ecol. Econ.* 126, 51–57. doi: 10.1016/j.ecolecon.2016.02.014

Soto, C. J. (2015). Is happiness good for your personality? concurrent and prospective relations of the big five with subjective well-being. *J. Personal.* 83, 45–55. doi: 10.1111/jopy.12081

Steptoe, A., Tsuda, A., and Tanaka, Y. (2007). Depressive symptoms, socio-economic background, sense of control, and cultural factors in university students from 23 countries. *Int. J. Behav. Med.* 14, 97–107. doi: 10.1007/BF03004175

Sucker, K., Both, R., Bischoff, M., Guski, R., Krämer, U., and Winneke, G. (2008). Odor frequency and odor annoyance Part II: dose–response associations and their modification by hedonic tone. *Int. Arch. Occup. Environ. Health* 81, 683–694. doi: 10.1007/s00420-007-0262-4

Sun, Q., and Yao, B. X. (2009). Study on security, interpersonal trust and the related factors among college students. *Soft Sci. Health.* 3, 290–293. doi: 10.3969/j.issn.1003-2800.2009.03.018

Taylor, S. E., and Brown, J. D. (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychol. Bull.* 103:193. doi: 10.1037/0033-2909.103.2.193

Tseng, K. K., Park, S. H., Shearston, J. A., Lee, L., and Weitzman, M. (2017). Parental psychological distress and family food insecurity: sad dads in hungry homes. *J. Dev. Behav. Pediatr.* 38, 611–618. doi: 10.1097/DBP.0000000000000481

Tynan, R. (2005). The effects of threat sensitivity and face giving on dyadic psychological safety and upward communication 1. *J. Appl. Soc. Psychol.* 35, 223–247. doi: 10.1111/j.1559-1816.2005.tb02119.x

Vail, J. (1999). "Insecure times: conceptualizing insecurity and security," in *Insecure Times: Living with Insecurity in Contemporary Society*, eds H. Michael, V. John, and W. Jane, (London: Routledge), 1–22.

Van der Wurff, A., Van Staalduinen, L., and Stringer, P. (1989). Fear of crime in residential environments: testing a social psychological model. *J. Soc. Psychol.* 129, 141–160. doi: 10.1080/00224545.1989.9711716

Van Hal, G. (2015). The true cost of the economic crisis on psychological well-being: a review. *Psychol. Res. Behav. Manag.* 8:17. doi: 10.2147/PRBM.S44732

Vieitez, J. C., Carcía, A. D. L. T., and Rodríguez, M. T. V. (2001). Perception of job security in a process of technological change: its influence on psychological well-being. *Behav. Inform. Technol.* 20, 213–223. doi: 10.1080/01449290120718

Weber, E. U., and Stern, P. C. (2011). Public understanding of climate change in the united states. *Am. Psychol.* 66, 315. doi: 10.1037/a0023253

Weinstein, N. D., Lyon, J. E., Rothman, A. J., and Cuite, C. L. (2000). Changes in perceived vulnerability following natural disaster. *J. Soc. Clin. Psychol.* 19, 372–395. doi: 10.1521/jscp.2000.19.3.372

Whitson, J. A., and Galinsky, A. D. (2008). Lacking control increases illusory pattern perception. *Science* 322, 115–117. doi: 10.1126/science.1159845

Williams, R. (2006). The psychosocial consequences for children and young people who are exposed to terrorism, war, conflict and natural disasters. *Curr. Opin. Psychiatry* 19, 337–349. doi: 10.1097/01.yco.0000228751.85828.c1

Wu, X., Yang, D. L., and Chen, L. (2017). The politics of quality-of-life issues: food safety and political trust in china. *J. Contemp. China* 26, 601–615. doi: 10.1080/10670564.2017.1274827

Xia, C., and Wei, T. (2011). Development of resident 's sense of security scale[J]. *China J. Health Psychol.* 19, 1126–1128. doi: 10.13342/j.cnki.cjhp.2011.09.021

Yan, Y. (2012). Food safety and social risk in contemporary china. *J. Asian Stud.* 71, 705–729. doi: 10.1017/S0021911812000678

Yin, P. P. (1980). Fear of crime among the elderly: some issues and suggestions. *Soc. Probl.* 27, 492–504. doi: 10.2307/800177

Yu, Q., and Zhao, Y. (2016). The influence of self objectification on appearance anxious among female college students—the mediating effect of security sense. *Adv. Psychol.* 6:452. doi: 10.12677/AP.2016.64060

Zani, B., Cicognani, E., and Albanesi, C. (2001). Adolescents' sense of community and feeling of unsafety in the urban environment. *J. Commun. Appl. Soc. Psychol.* 11, 475–489. doi: 10.1002/casp.647

Zenker, S., and Petersen, S. (2014). An integrative theoretical model for improving resident-city identification. *Environ. Plan. A* 46, 715–729. doi: 10.1068/a46191

Zhang, F. (2009). The relationship between state attachment security and daily interpersonal experience. *J. Res. Personal.* 43, 511–515. doi: 10.1016/j.jrp.2008.12.026

Zhang, H., Chan, D. K. S., Teng, F., and Zhang, D. (2015). Sense of interpersonal security and preference for harsh actions against others: the role of dehumanization. *J. Exp. Soc. Psychol.* 56, 165–171. doi: 10.1016/j.jesp.2014.09.014

Zhang, X., Chen, X., and Zhang, X. (2018). The impact of exposure to air pollution on cognitive performance. *Proc. Natl. Acad. Sci. U.S.A.* 115, 9193–9197. doi: 10.1073/pnas.1809474115

Zhang, Y. (2007). Synthetic index of sense of security of the Residents in Beijing. *J. Capital Univ. Econ. Bus.* 20, 115–117. doi: 10.3969/j.issn.1008-2700.2007.02.021

Zhao, J., and Jing, F. (2013). Antecedents and effects of extensive familism consciousness in online brand community. *Bus. Rev.* 27, 88–98. doi: 10.14120/j.cnki.cn11-5057/f.2015.12.009

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 06 December 2019
doi: 10.3389/fpsyg.2019.02771

Exploratory and Confirmatory Factor Analysis of the 9-Item Utrecht Work Engagement Scale in a Multi-Occupational Female Sample: A Cross-Sectional Study

Mikaela Willmer[1], Josefin Westerberg Jacobson[1,2] and Magnus Lindberg[1,2]*

[1] Department of Health and Caring Sciences, Faculty of Health and Occupational Studies, University of Gävle, Gävle, Sweden, [2] Department of Public Health and Caring Sciences, Uppsala University, Uppsala, Sweden

OPEN ACCESS

Edited by:
Elisa Pedroli,
Italian Auxological Institute (IRCCS), Italy

Reviewed by:
Silvia Testa,
University of Turin, Italy
István Tóth-Király,
Concordia University, Canada

****Correspondence:***
Mikaela Willmer
Mikaela.Willmer@hig.se

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *10 April 2019*
Accepted: *25 November 2019*
Published: *06 December 2019*

Citation:
Willmer M, Westerberg Jacobson J and Lindberg M (2019) Exploratory and Confirmatory Factor Analysis of the 9-Item Utrecht Work Engagement Scale in a Multi-Occupational Female Sample: A Cross-Sectional Study. Front. Psychol. 10:2771. doi: 10.3389/fpsyg.2019.02771

Objective: The aim of the present study was to use exploratory and confirmatory factor analysis (CFA) to investigate the factorial structure of the 9-item Utrecht work engagement scale (UWES-9) in a multi-occupational female sample.

Methods: A total of 702 women, originally recruited as a general population of 7–15-year-old girls in 1995 for a longitudinal study, completed the UWES-9. Exploratory factor analysis (EFA) was performed on half the sample, and CFA on the other half.

Results: Exploratory factor analysis showed that a one-factor structure best fit the data. CFA with three different models (one-factor, two-factor, and three-factor) was then conducted. Goodness-of-fit statistics showed poor fit for all three models, with RMSEA never going lower than 0.166.

Conclusion: Despite indication from exploratory factor analysis (EFA) that a one-factor structure seemed to fit the data, we were unable to find good model fit for a one-, two-, or three-factor model using CFA. As previous studies have also failed to reach conclusive results on the optimal factor structure for the UWES-9, further research is needed in order to disentangle the possible effects of gender, nationality and occupation on work engagement.

Keywords: confirmatory factor analysis, exploratory factor analysis, Utrecht work engagement scale, work engagement, occupational psychology

INTRODUCTION

Work engagement has been described as the conceptual opposite of burnout (González-Romá et al., 2006), and as such belongs in the area of positive psychology, or "the study of the conditions and processes that contribute to the flourishing or optimal functioning of people, groups, and institutions"(Gable and Haidt, 2005). In occupational health, the study of work engagement focuses on factors that contribute to job satisfaction as well as long-term mental and physical health (Torp et al., 2013).

Work engagement has been described as "a positive work-related state of mind characterized by vigor, dedication and absorption." (Schaufeli et al., 2002). These three concepts are in their turn described as "characterized by high levels of energy and mental resilience while working, the willingness to invest effort in one's work, and persistence even in the face of difficulties" (Vigor), "characterized by a sense of significance, enthusiasm, inspiration, pride and challenge" (Dedication) and "characterized by being fully engrossed in one's work, so that time passes quickly and one has difficulties in detaching oneself from work" (Absorption) (Schaufeli et al., 2002).

The idea that these three concepts – Vigor, Dedication and Absorption – together form the foundation of work engagement forms the basis of the Utrecht work engagement scale (UWES) (Schaufeli et al., 2002). Originally a 17-item questionnaire (UWES-17), the original authors have shortened it to a 9-item version (UWES-9) in order to reduce the burden on the respondents and minimize attrition (Schaufeli et al., 2006). The items are in the form of statements (for example "At my work, I feel bursting with energy" (Vigor); "I find the work that I do full of meaning and purpose" (Dedication); "When I am working, I forget everything else around me" (Absorption) which the respondent reads and reacts to by indicating one of 7 points on a scale ranging from 0 ("Never") to 6 ("All the time"). The 9-item version, which has been psychometrically tested in various countries and samples (Ho Kim et al., 2017; Petrović et al., 2017), will be the focus of the present study.

In a number of studies, conducted in different countries and with samples of various make-ups, UWES-9 scores have been found to be associated with work performance, job satisfaction, and mental and physical health (Bakker and Matthijs Bal, 2010; Christian et al., 2011). The scores have also been found to predict general life satisfaction and the frequency of sickness absence (Leijten et al., 2015).

Despite its wide-spread use, both the UWES-17 and the UWES-9 have been the subject of some criticism. Mills et al. (2012) have argued that the methodology when developing the original scale contained flaws in relation to the establishment of its factorial structure. Criticism has also been voiced regarding the factor structure of the instrument, one of the main points being that the three subscales Vigor, Dedication and Absorption are very closely correlated with each other, casting doubt on the three-factor structure's superiority to a one-factor structure using only the total score on the scale (Kulikowski, 2017). For example, Shirom has argued that the three dimensions of Vigor, Dedication, and Absorption were not theoretically deduced and that they overlap each other conceptually (Shirom, 2003). In support of this, several studies have failed to confirm the three-factor structure in their samples. Previous studies have also tested other factor structures – for example, Kulikowski (2019) tested a two-factor structure, with Dedication and Vigor merged into a single factor and Absorption constituting a second factor (Kulikowski, 2019). A 2017 review by Kulikowski investigated the factorial structure of the UWES-17 and UWES-9 as reported in 21 different studies, conducted in 24 countries using samples from a variety of occupations and countries. The author found that of the 11 studies investigating the UWES-9, three confirmed the one-factor structure, three the three-factor structure, four studies found these two factor structures to be equivalent, and one study failed to support either alternative (Kulikowski, 2017). Thus, Kulikowski (2017) concluded that no definitive recommendations could be made based on the review. He also pointed out the importance, in light of these inconclusive results, that further research be conducted on the factorial structure of the UWES-9 in different samples (Kulikowski, 2017).

Only one previous study has tested the factorial validity of the UWES-9 in a Swedish sample (Hallberg and Schaufeli, 2006). In their sample of 186 information communication technology consultants (of whom 37% were women), both the one-factor and three-factor structures were supported by data, leading the authors to draw the conclusion that both options were equally strong. If the scope is broadened to take in all the Scandinavian countries, a Norwegian study using a large multi-occupational sample (n = 1266, 67% women) found support for the three-factor structure, but also found that the three latent factors were strongly correlated, leading the authors to suggest that a one-factor structure might also be suitable (Nerstad, Richardsen and Martinussen, 2010). In addition to this, a Finnish study found, in a sample of 9404 workers in several different occupational sectors, that both the one-factor and three-factor structures may reasonably be used (Seppälä et al., 2009). Similarly to the Norwegian study, the results showed that the three subscales of Vigor, Dedication, and Absorption were highly correlated.

Interestingly, it has been suggested that as a rule, levels of work engagement tend to be higher in countries in Northwestern Europe, and lower in Southern Europe, on the Balkans and in Turkey (Schaufeli, 2018). However, Sweden is identified as an exception to this rule, with relatively low levels of work engagement compared to, for example, Norway, where levels were found to be higher (Schaufeli, 2018).

The 9-item UWES is a widely used instrument to measure work engagement. Despite this, the optimal factorial structure of the UWES-9 remains unknown. A recent review of factorial structure for the UWES-9 and UWES-17 failed to reach conclusive results, and indicated that more research was needed to determine the appropriate default factorial structure (Kulikowski, 2017). Many previous studies have used relatively small samples, and many have reached inconclusive results, including the only previously published Swedish study. In order to adequately assess and potentially target work engagement in future interventions using Swedish populations, it is important to examine and ascertain whether Swedish people hold the same representation of work engagement. Thus, the aim of the present study was to use exploratory and confirmatory factor analysis (CFA) to investigate the factorial structure of the 9-item UWES in a multi-occupational Swedish sample.

MATERIALS AND METHODS

Participants

The women in the all-female sample used for the current study were originally recruited in 1995, when they were aged between 7 and 15 years, through stratified randomization from a number of school classes in Sweden. They were sampled to represent a general population of girls, and were participants in a longitudinal study aiming to identify risk and protective factors for the development of eating disorders. More details about the recruitment and follow-up can be found elsewhere (Westerberg-Jacobson et al., 2010). The data used in the current study was collected in 2015, as part of the 20-year follow-up data collection. The participants remaining in the study were asked to complete a number of questionnaires, including the UWES-9, and those who indicated that they were currently working full-time or part-time (not on long-term sick-leave, parental leave, unemployed, or studying full-time) were included in the current study. Thus, the final sample consisted of 702 women, aged between 26 and 37, who completed a Swedish translation of the 9-item UWES (Schaufeli et al., 2006). Aside from the UWES-9, data was collected on level of education (primary school, secondary education or university education), although not on specific occupation.

Ethics Statement

The project was approved by the Regional Ethics Board in Uppsala, Sweden (2014/401). At the time of the original recruitment, in 1995, the participants and their parents gave written informed consent to take part in the study. At the time of the data collection for the present study, the participants again gave their written informed consent and were reminded that their participation was voluntary, could be withdrawn any time without giving a reason, and that all information would be treated confidentially. All participants who completed the data collection were offered a cinema ticket or a department store gift voucher as thanks.

Statistical Analysis

All analyses were performed using Stata 14 (StataCorp, 2015) and SPSS (IBM Corp, 2016) statistical software packages. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy and Bartlett's Test of Sphericity were used to assess the suitability of the data for factor analysis (Dziuban and Shirkey, 1974). Exploratory factor analysis (EFA) was first performed unrotated, using maximum likelihood extraction and eigenvalues > 1. Additionally, we performed EFA with promax rotation and enforcing three-factor solution in order to test the theoretical structure of the UWES-9. In this analysis, we also used maximum likelihood extraction. Additionally, Parallel Analysis (using principal axis factoring) and Velicer's Minimum Average Partial test were conducted (O'Connor, 2000).

CFA was then performed using maximum likelihood estimation.

In order to investigate the models' goodness of fit, a number of statistics were used: Overall χ^2 (Hooper et al., 2008), root mean square error of approximation (RMSEA) (Steiger, 1990; Hooper et al., 2008), Akaike's information criterion (AIC), Bayesian information criterion (BIC), comparative fit index (CFI), Tucker-lewis index (TLI) (Bentler, 1990), and the standardized root mean square residual (SRMSR) (Hooper et al., 2008).

RESULTS

Demographic information about the participants can be seen in **Table 1**. Data on highest attained educational level was collected, and showed that the majority of the sample had attended at least 3 years of higher education.

TABLE 1 | Demographic information about the participants.

Variable (*n* = 702)		
	Mean	**Standard deviation**
Age	31.8 (2.9)	
Marital status	**Frequency**	**Percentage**
Single	159	23
Married/cohabiting	530	76
Divorced	9	1
Education	**Frequency**	**Percentage**
Compulsory (9 years)	9	1
<3 years upper secondary	21	3
≥3 years upper secondary	152	22
<2 years university	75	11
≥2 years university	425	61
UWES scores	**Mean**	**Standard deviation**
Total UWES score	4.06	1.18
Vigor	3.96	1.19
Dedication	4.24	1.25
Absorption	3.98	1.32

The inter-item correlation was relatively high for all items of the UWES-9, ranging between 0.524 and 0.849. The three subscales Vigor (V), Dedication (D), and Absorption (A) also showed high correlation with each other (0.79–0.84). In addition to this, Cronbach's alpha was calculated and found to be 0.947, indicating very good internal consistency.

The items were checked for skewness and kurtosis and these are shown in **Table 2**, together with the wording of the items, their respective subscales, mean scores and standard deviations. Based on the Shapiro-Wilks test and a visual inspection of their histograms, normal Q-Q plots and box-plots, we concluded that the UWES item distributions had a skewness range between −0.560 and −1.262 (SE = 0.094) and a kurtosis range between −0.046 and 1.645 (SE = 0.187) (**Table 2**). The values for skewness and kurtosis were deemed to be within the range for maximum likelihood estimation. We also tested the multivariate normality using Doornik-Hansen test, the Mardia skewness test and Mardia kurtosis test. For all of these, the *p*-value was <0.0001, indicating non-normality.

In the next step, the sample was randomly divided in two, so that mutually independent samples were obtained for the EFA and CFA, respectively. As the number of participants with missing values was very low (19 individuals, corresponding to 3% of the entire sample), only observations without any missing items were used, resulting in 683 observations in total, 341 for the EFA and 342 for the CFA.

Exploratory Factor Analysis

The results of the EFA suggested that one factor explained over 70% of the variance. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was 0.922, indicating that the sample was

TABLE 2 | Items with their subscales, mean scores, standard deviations, skewness, and kurtosis.

Item (subscale)	Mean	Standard deviation	Skewness	Kurtosis
1. At my work, I feel bursting with energy (V)	3.93	1.30	−0.798	0.294
2. At my job, I feel strong and vigorous (V)	4.08	1.22	−0.921	0.678
3. I am enthusiastic about my job (D)	4.10	1.32	−0.900	0.568
4. My job inspires me (D)	4.01	1.44	−0.808	0.266
5. When I get up in the morning, I feel like going to work (V)	3.89	1.49	−0.805	0.113
6. I feel happy when I am working intensely (A)	3.89	1.49	−0.711	−0.046
7. I am proud of the work that I do (D)	4.62	1.32	−1.262	1.645
8. I get carried away when I am working (A)	4.43	1.31	−1.159	1.474
9. I am immersed in my work (A)	3.64	1.64	−0.560	−0.497

V, vigor; D, dedication; A, absorption.

TABLE 3 | Factor loadings.

Variable	Factor 1
UWES1	0.78
UWES2	0.81
UWES3	0.93
UWES4	0.90
UWES5	0.81
UWES6	0.86
UWES7	0.78
UWES8	0.79
UWES9	0.65

adequate, and Bartlett's Test of Sphericity gave a p-value of <0.001. A Scree plot of the eigenvalues was constructed (not shown) and shown to be strongly in favor of the one-factor structure. The χ2 for this model was 332,43 (df 27).

Velicer's MAP test was also performed, both in the original (Velicer, 1976) and revised version (O'Connor, 2000). This also strongly pointed toward a one-factor solution.

Finally, in the Parallel Analysis, the raw data eigenvalue from the actual data was greater than eigenvalues of the 95th percentile of the distribution of random data for four factors, in disagreement with the MAP test and the EFA (O'Connor, 2000).

Table 3 shows the factor loadings. As the table shows, all loadings were relatively high, ranging from 0.65 to 0.93.

In addition to this, we also conducted EFA using promax rotation and enforcing a three-factor structure, in order to compare the fit of the theoretical dimensionality of the UWES-9 with the one-factor solution we found in our sample. The χ2 for this model was 45,72 (df 12) ($p < 0.001$). The items did not load on their expected factors "Dedication" had 4 items (3, 4, 5, 6), "Vigor" had 2 items (1, 2), and "Absorption" had 3 items (7, 8, 9).

Confirmatory Factor Analysis

As the EFA suggested a one-factor solution, as described above, the model was first specified with just one latent factor (Work Engagement). Standardized coefficients were used and the estimation model was maximum likelihood, since the items showed acceptable skewness and kurtosis (**Table 2**). Observations with missing values were excluded.

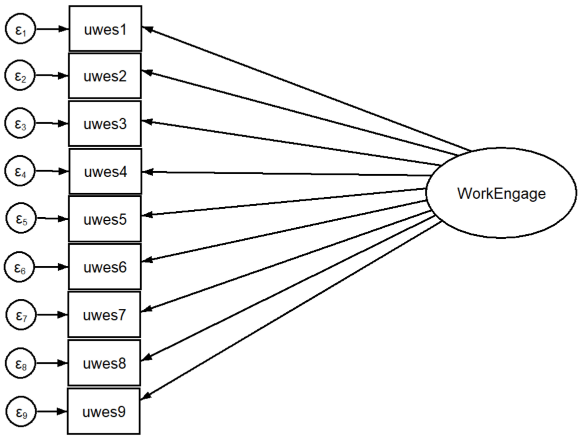


FIGURE 1 | One-factor structure with maximum likelihood estimation.

In order to also test the theoretical foundation of the UWES-9, we performed CFA with the original three subscales Vigor, Dedication and Absorption. Additionally, inspired by a previous study by Kulikowski (2019), who also tested a two-factor model, we also performed CFA using this structure.

Figures 1–3 show all the attempted models.

Table 4 shows the coefficients of the hypothesized relationships, together with their z-values, standard errors, 95% confidence intervals and p-values, for all tested models.

After estimating the models, goodness-of-fit statistics were obtained, as described in the section "Materials and Methods," above. As can be seen in **Table 5**, none of the models showed very good fit, with RMSEA ranging between 0.181 and 0.167. Also, CFI and TLI, which should preferably be above 0.95 (Hooper et al., 2008) remained below this value for all tested models.

DISCUSSION

The aim of the present study was to use exploratory and CFA to investigate the factorial structure of the UWES in a multi-occupational sample of Swedish women. The EFA seemed to mainly favor a one-factor solution, which was shown to explain over 70% of the variance.

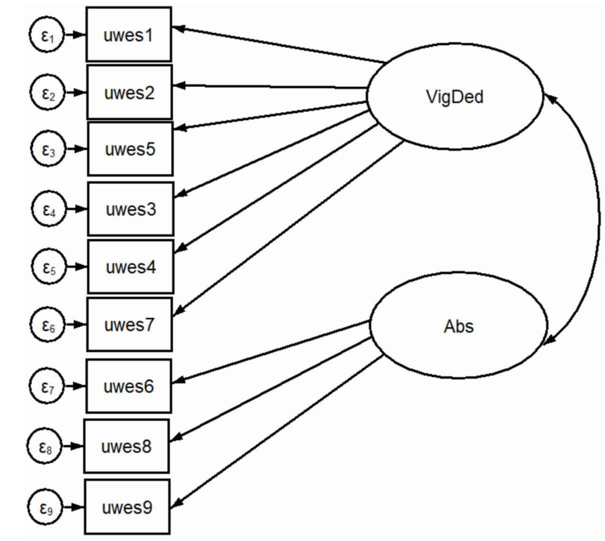


FIGURE 2 | Two-factor structure with maximum likelihood estimation.

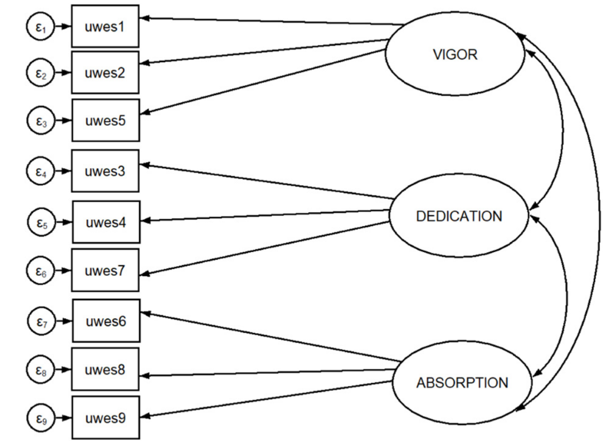


FIGURE 3 | Three-factor structure with maximum likelihood estimation.

TABLE 4 | All models' standardized coefficients and associated data.

Item	Coefficient	Standard error	*z*-value	*p*-value	95% CI
One-factor model					
Item 1	0.79	0.02	50.42	<0.0001	0.76; 0.82
Item 2	0.82	0.01	59.95	<0.0001	0.79; 0.85
Item 3	0.92	0.01	132.99	<0.0001	0.91; 0.94
Item 4	0.90	0.01	109.15	<0.0001	0.89; 0.92
Item 5	0.81	0.01	55.85	<0.0001	0.78; 0.83
Item 6	0.87	0.01	83.55	<0.0001	0.85; 0.89
Item 7	0.76	0.02	44.83	<0.0001	0.73; 0.80
Item 8	0.81	0.01	57.54	<0.0001	0.78; 0.84
Item9	0.69	0.02	33.19	<0.0001	0.65; 0.73
Two-factor model*					
Item 1	0.80	0.02	36.08	<0.0001	0.75; 0.84
Item 2	0.83	0.02	42.84	<0.0001	0.79; 0.87
Item 3	0.92	0.01	80.72	<0.0001	0.89; 0.94
Item 4	0.90	0.01	67.82	<0.0001	0.87; 0.92
Item 5	0.76	0.02	31.74	<0.0001	0.72; 0.81
Item 6	0.89	0.02	56.27	<0.0001	0.86; 0.92
Item 7	0.77	0.02	32.23	<0.0001	0.72; 0.81
Item 8	0.83	0.02	40.16	<0.0001	0.79; 0.87
Item 9	0.76	0.03	28.72	<0.0001	0.71; 0.81
Three-factor model**					
Item 1	0.89	0.01	81.70	<0.0001	0.87; 0.91
Item 2	0.92	0.01	93.13	<0.0001	0.90; 0.94
Item 3	0.94	0.01	147.03	<0.0001	0.93; 0.95
Item 4	0.93	0.01	128.89	<0.0001	0.91; 0.94
Item 5	0.74	0.02	36.26	<0.0001	0.70; 0.78
Item 6	0.99	0.01	80.71	<0.0001	0.86; 0.90
Item 7	0.75	0.02	42.25	<0.0001	0.72; 0.79
Item 8	0.84	0.02	60.96	<0.0001	0.81; 0.86
Item 9	0.73	0.02	36.10	<0.0001	0.69; 0.77

Items 1, 2, 3, 4, 5, and 7 belong to the combined vigor/dedication factor. Items 6, 8, and 9 belong to the absorption factor. Items 1, 2, and 4 belong to the vigor factor. Items 3, 4, and 7 belong to the dedication factor. Items 6, 8, and 9 belong to the absorption factor.*

TABLE 5 | Goodness-of-fit statistics for all models.

Fit statistic	One-factor model	Two-factor model	Three-factor model
Chi2 (df)	633.90 (27)	354.49 (26)	247.76 (24)
RMSEA (90% CI)	0.181 (0.169; 0.194)	0.192 (0.175; 0.192)	0.167 (0.154; 0.180)
AIC	16221.47	8246.29	8143.56
BIC	16343.70	8353.66	8258.60
CFI	0.895	0.882	0.920
TLI	0.860	0.837	0.880
SRMR	0.046	0.049	0.065

Df, degrees of freedom; RMSEA, root mean squared error of approximation; CI, confidence interval; AIC, Akaike's information criterion; BIC, Bayesian information criterion; CFI, comparative fit index; TLI, Tucker-Lewis index; SRMR, standardized root mean squared residual.

Confirmatory factor analysis was then performed using three different models: one-factor, two-factor, and three-factor. Goodness-of-fit statistics were obtained for all models and showed that none of them showed overall good fit, with RMSEA never going below 0.167 and CFI and TLI remaining relatively low (**Table 5**).

As previously mentioned, a recent review of the factorial structure of the UWES showed inconclusive results, with some included studies showing best fit for a one-factor structure, some showing best fit for a three-factor structure, and some showing an equally good (or poor) fit for both (Kulikowski, 2017). This indicates a need for further research into the underlying factors impacting the factor structures in various samples.

One of the studies included in the Kulikowski review found that neither the one-factor nor the three-factor structure of the UWES-9 was a good fit for their data (Wefald et al., 2012). This used a sample similar to ours, both in terms of size (382 vs. 342) and level of education (in both samples, around 60% had a university degree or higher). The RMSEA was 0.18 and 0.16 for the one-factor and three-factor structures, in the Wefald study, almost identical to 0.181 and 0.167 for our study.

A previous study by Kulikowski (2019) has also attempted a two-factor structure, merging Dedication and Vigor into a single factor, letting Absorption constitute the second factor (Kulikowski, 2019). We attempted the same model in the present study, but in agreement with Kulikowski's results, failed to obtain satisfactory goodness of fit.

The only previous Swedish study using the UWES used a sample consisting of 186 information technology (IT) consultants (37% women) and found that both the one-factor and three-factor structure showed similar fit, with RMSEA of 0.13 and CFI of 0.97 for both (Hallberg and Schaufeli, 2006). Although this sample was Swedish, it was different from that of the present study in other significant ways, such as gender (a majority were male) and occupation (all the participants were IT consultants, whilst ours was a multi-occupational sample), which may explain the differences in the results.

If our results are compared with those of other studies also using multi-occupational samples, several of them have, in agreement the Swedish study by Hallberg and Schaufeli (2006), found that *both* the one-factor and three-factor structures may be used. For example, this was the case for Schaufeli et al. (2006) with a very large multinational sample of 14521 individuals.

These differing results support the recommendation made by Kulikowski (2017), namely that each study using the UWES-9 should undertake their own factor analysis based on their own sample, and make a decision on which structure to use based on their own results (Kulikowski, 2017). In addition to this, and in agreement with the current study, several previous studies have found that none of the factor structures tested have shown an acceptable fit (Hallberg and Schaufeli, 2006; Wefald et al., 2012). Subsequently, researchers looking to use a measure of work engagement may wish to use another instrument in parallel with the UWES.

The present study has strengths, as well as weaknesses. The relatively large sample size of approximately 700 women made it possible to randomly divide the group into half so that both an exploratory and a CFA could be undertaken. The fact that the sample consisted exclusively of women may be seen both as a strength and as a weakness. On the one hand, it ensures that the results are not skewed by an uneven gender balance, but on the other hand our results should not be assumed to be generalizable to males. An Iranian study investigating determinants of work engagement in hospital staff found no significant effect of gender (Mahboubi et al., 2014). However, a Dutch study exploring work engagement and burnout in veterinarians found that women rated their work engagement lower than men, indicating that gender differences may vary with different occupational groups, nationalities, or other, hitherto unknown factors (Mastenbroek et al., 2014).

In addition to this, in terms of generalizability, it should be acknowledged that the sample used in the present study should be considered to represent the white-collar population, based on the higher-than-average level of education. More than 60% of the participants reported having at least 3 years of university education, whilst the national average for women between the ages of 25 and 34 is 35%, according to Statistics Sweden (Statistics Sweden, 2017). In addition to this, only Swedish-speaking girls participated. However, 21.6% had immigrated or had parents who had immigrated to Sweden, which is in line with the population in general (Statistics Sweden, 2018).

CONCLUSION

The present study used a large, multi-occupational female sample to explore the factorial structure of the UWES-9. Despite indication from EFA that a one-factor structure best fit the data, we were unable to find good model fit for a one-, two-, or three-factor model using CFA. As previous studies have also failed to reach conclusive results on the optimal factor structure for the UWES-9, further research is needed in order to disentangle the possible effects of gender, nationality and occupation on work engagement. Until such data exists, researchers would be wise to conduct their own factor analysis in order to determine whether the total score, the three dimensions representing Vigor, Dedication and Absorption, or even a two-factor structure is applicable for their sample.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This project was approved by the Regional Ethics Board (2014/401). At the time of the data collection for the present study, the participants were again asked to give their consent and reminded that their participation was voluntary, could be withdrawn any time without giving a reason, and that all information would be treated confidentially. All participants who completed the data collection were offered a cinema ticket or a department store gift voucher as thanks.

AUTHOR CONTRIBUTIONS

MW contributed to the conception and design of the work, performed the analyses, and drafted the manuscript. JW and ML contributed to the conception and design of the work, took part in the data collection and analyses, and revised the work critically. All authors approved the final version to be published, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

FUNDING

This work was supported by Capio Research Foundation, the Signe and Olof Wallenius Foundation, and the Thuring Foundation.

REFERENCES

Bakker, A. B., and Matthijs Bal, P. (2010). Weekly work engagement and performance: a study among starting teachers. *J. Occup. Organ. Psychol.* 83, 189–206. doi: 10.1348/096317909X402596

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Bollen, K. A. (2014). *Structural Equations with Latent Variables*. New York, NY: Wiley.

Christian, M. S., Adela, S. G., and Slaughter, J. E. (2011). Work engagement: a quantitative review and test of its relation with task and contextual performance. *Pers. Psychol.* 64, 89–136. doi: 10.1111/j.1744-6570.2010.01203.x

IBM Corp (2016). *SPSS for Windows*. Armonk, NY: IBM Corp.

Dziuban, C. D., and Shirkey, E. C. (1974). When is a correlation matrix appropriate for factor analysis? Some decision rules. *Psychol. Bull.* 81, 358–361. doi: 10.1037/h0036316

Gable, S. L., and Haidt, J. (2005). What (and why) is positive psychology? *Rev. Gen. Psychol.* 9, 103–110. doi: 10.1037/1089-2680.9.2.103

González-Romá, V., Schaufeli, W. B., Bakker, A. B., and Lloret, S. (2006). Burnout and work engagement: independent factors or opposite poles? *J. Vocat. Behav.* 68, 165–174. doi: 10.1016/j.jvb.2005.01.003

Hallberg, U. E., and Schaufeli, W. B. (2006). "Same same" but different? Can work engagement be discriminated from job involvement and organizational commitment? *Eur. Psychol.* 11, 119–127. doi: 10.1027/1016-9040.11.2.119

Ho Kim, W., Park, J. G., and Kwon, B. (2017). Work engagement in South Korea. *Psychol. Rep.* 120, 561–578. doi: 10.1177/0033294117697085

Hooper, D., Coughlan, J., and Mullen, M. (2008). Structural equation modelling: guidelines for determining model fit. *Electron. J. Bus. Res. Methods* 6, 53–60.

Kulikowski, K. (2017). Do we all agree on how to measure work engagement? Factorial validity of Utrecht work engagement scale as a standard measurement tool – A literature review. *Int. J. Occup. Med. Environ. Health* 30, 161–175. doi: 10.13075/ijomeh.1896.00947

Kulikowski, K. (2019). One, two or three dimensions of work engagement? Testing the factorial validity of the Utrecht work engagement scale on a sample of Polish employees. *Int. J. Occup. Saf. Ergon.* 25, 241–249. doi: 10.1080/10803548.2017.1371958

Leijten, F., van den Heuvel, S. G., van der Beek, A. J., Ybema, J. F., Robroek, S. J., and Burdorf, A. (2015). 'Associations of work-related factors and work engagement with mental and physical health: a 1-year follow-up study among older workers. *J. Occup. Rehabil.* 25, 86–95. doi: 10.1007/s10926-014-9525-6

Mahboubi, M., Ghahramani, F., Mohammadi, M., Amani, N., Mousavi, S. H., Moradi, F., et al. (2014). Evaluation of work engagement and its determinants in Kermanshah hospitals staff in 2013. *Glob. J. Health Sci.* 7, 170–176. doi: 10.5539/gjhs.v7n2p170

Mastenbroek, N. J., Jaarsma, A. D., Demerouti, E., Muijtjens, A. M., Scherpbier, A. J., and van Beukelen, P. (2014). Burnout and engagement, and its predictors in young veterinary professionals: the influence of gender. *Vet. Rec.* 174:144. doi: 10.1136/vr.101762

Mills, M., Culbertson, S., and Fullagar, C. (2012). Conceptualizing and measuring engagement: an analysis of the Utrecht work engagement scale. *J. Happiness Stud.* 13, 519–545. doi: 10.1007/s10902-011-9277-3

Nerstad, C. G. L., Richardsen, A. M., and Martinussen, M. (2010). Factorial validity of the Utrecht work engagement scale (UWES) across occupational groups in Norway. *Scand. J. Psychol.* 51, 326–333. doi: 10.1111/j.1467-9450.2009.00770.x

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behav. Res. Methods Instrum. Comput.* 32, 396–402. doi: 10.3758/bf03200807

Petrović, I. B., Vukelić, M., and čizmić, S. (2017). Work engagement in Serbia: psychometric properties of the Serbian version of the Utrecht work engagement scale (UWES). *Front. Psychol.* 8:1799. doi: 10.3389/fpsyg.2017.01799

Schaufeli, W. (2018). Work engagement in Europe: relations with national economy, governance and culture. *Organ. Dyn.* 47, 99–106.

Schaufeli, W. B., Bakker, A. B., and Salanova, M. (2006). The measurement of work engagement with a short questionnaire a cross-national study. *Educ. Psychol. Meas.* 66, 701–716. doi: 10.1177/0013164405282471

Schaufeli, W. B., Salanova, M., González-romá, V., and Bakker, A. B. (2002). The measurement of engagement and burnout: a two sample confirmatory factor analytic approach. *J. Happiness Stud.* 3, 71–92.

Seppälä, P., Mauno, S., Feldt, T., Hakanen, J., Kinnunen, U., Tolvanen, A., et al. (2009). The construct validity of the Utrecht work engagement scale: multisample and longitudinal evidence. *J. Happiness Stud.* 10, 459–481. doi: 10.1007/s10902-008-9100-y

Shirom, A. (2003). "Feeling vigorous at work? The construct of vigor and the study of positive affect in organizations," in *Emotional and Physiological Processes and Positive Intervention Strategies (Research in Occupational Stress and Well-being*, Vol. 3, eds P. L. Perrewe, and D. C. Ganster, (Bingley: Emerald Group Publishing Limited), 135–164. doi: 10.1016/s1479-3555(03)03004-x

StataCorp (2015). *Stata Statistical Software: Release 14*. College Station, TX: StataCorp.

Statistics Sweden (2017). *Befolkningens Utbildning*. Available at: http://www.scb.se/uf0506 (accessed July 2, 2018).

Statistics Sweden (2018). *Folkmängd Och Befolkningsförändringar 2017*. Available at: http://www.scb.se/hitta-statistik/statistik-efter-amne/befolkning/befolkningens-sammansattning/befolkningsstatistik/pong/statistiknyhet/folkmangd-och-befolkningsforandringar-20172/ (accessed July 2, 2018).

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behav. Res.* 25, 173–180. doi: 10.1207/s15327906mbr2502_4

Torp, S., Grimsmo, A., Hagen, S., Duran, A., and Gudbergsson, S. B. (2013). Work engagement: a practical measure for workplace health promotion? *Health Promot. Int.* 28, 387–396. doi: 10.1093/heapro/das022

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika* 41, 321–327. doi: 10.1007/bf02293557

Wefald, A. J., Mills, M. J., Smith, M. R., and Downey, R. G. (2012). A comparison of three job engagement measures: examining their factorial and criterion-related validity. *Appl. Psychol. Health Well Being* 4, 67–90. doi: 10.1111/j.1758-0854.2011.01059.x

Westerberg-Jacobson, J., Edlund, B., and Ghaderi, A. (2010). A 5-year longitudinal study of the relationship between the wish to be thinner, lifestyle behaviours and disturbed eating in 9-20-year old girls. *Eur. Eat. Disord. Rev.* 18, 207–219. doi: 10.1002/erv.98

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 January 2020
doi: 10.3389/fpsyg.2019.02903

Evaluating the Dimensionality and Psychometric Properties of the Brief Self-Control Scale Amongst Chinese University Students

Sai-fu Fung[1], Chris Yiu Wah Kong[1] and Qian Huang[2]*

[1] Department of Social and Behavioural Sciences, City University of Hong Kong, Kowloon, Hong Kong, [2] Department of Sports Training, Xi'an Physical Education University, Xi'an, China

The aim of this study was to assess the dimensionality and psychometric properties of the Brief Self-Control Scale (BSCS) using a sample of university students in mainland China. Nine hundred and three students from a Chinese university participated in this study. The internal consistency, criterion validity, factorial validity and construct validity of the scale were examined. The Chinese versions of the BSCS demonstrated good internal consistency with a Cronbach's alpha of 0.81. The BSCS also showed significant moderate correlations with other construct-related scales. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) suggested that only a modified 11-item BSCS with a four-factor structure was a good model fit in the sample of Chinese university students, as χ^2 (106.626)/37 = 2.88, SRMR = 0.036, comparative fit index (CFI) = 0.992, Tucker-Lewis fit index (TLI) = 0.989, RMSEA = 0.046. The implications for research and theoretical development are discussed.

Keywords: Brief Self-Control Scale, Chinese, confirmatory factor analysis, personality, self-control, university students, validation

OPEN ACCESS

Edited by:
Elisa Pedroli,
Italian Auxological Institute (IRCCS),
Italy

Reviewed by:
Edson Filho,
University of Central Lancashire,
United Kingdom
Shuqiao Yao,
Central South University, China

****Correspondence:***
Sai-fu Fung
sffung@cityu.edu.hk

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *22 February 2019*
Accepted: *06 December 2019*
Published: *08 January 2020*

Citation:
Fung S, Kong CYW and Huang Q
(2020) Evaluating the Dimensionality
and Psychometric Properties of the
Brief Self-Control Scale Amongst
Chinese University Students.
Front. Psychol. 10:2903.
doi: 10.3389/fpsyg.2019.02903

INTRODUCTION

Since the inception of impulse control and self-control concepts in the early 70s there has been extensive empirical research on their psychometric properties, theoretical underpinnings, and behavioral implications (Mischel, 1974; Ainslie, 1975). Many scholars regard self-control as essential for human positive growth and development (Metcalfe and Mischel, 1999; Tangney et al., 2004; Duckworth and Kern, 2011; de Ridder et al., 2012). Twentieth-century measurements of self-control, such as the self-control rating scale (Kendall and Wilcox, 1979), the bonding self-control scale (SCS) (Gottfredson, 1990), and Grasmick's SCS (Grasmick et al., 1993), were commonly used for criminological and addictive studies amongst children and juvenile delinquents. These scales were evaluated and applied to different criminological research projects involving children and juveniles (Wang, 2002; Piquero and Bouffard, 2007; Weng and Chui, 2018). Studies suggest that whilst people with higher self-control are inclined to delay gratification and are high achievers, those with lower self-control are less likely to inhibit impulsive behavior (Mischel and Mischel, 1983; Baumeister, 2016). SCSs have been used to analyse the relationship between emotional exhaustion and counterproductive workplace behaviors. In particular, Maloney et al. (2012) found that impulsivity was positively and significantly related to both interpersonally directed and organizationally directed counterproductive workplace behaviors, whereas restraint was negatively related to emotional exhaustion when controlling for the effects of impulsivity. Research also suggests that self-control is an important risk and protective factor amongst jail inmates (Malouf et al., 2014).

In the literature on personality, self-control has been recently associated with positive psychological adjustment and a broad range of positive outcomes in life, such as happiness, well-being and quality of life (Rothbaum et al., 1982; Tangney, 1991, 1995; Baumeister, 1994; Tangney et al., 1996; Eisenberg et al., 1998; Fabes et al., 1999). As such, Tangney et al. (2004) had developed the 36-item SCS and the shortened 13-item Brief Self-Control Scale (BSCS). The development and validation of these two scales signifies that self-control concepts can be more scientifically applied to various types of performance such as academic attainment, the formation of good habits, refraining from distractions and controlling of urges and impulsive behavior such as procrastination and drug-taking.

Brief Self-Control Scale has been translated into different languages and validated by the French-speaking population of Canada (Brevers et al., 2017), and in Germany (Bertrams and Dickhäuser, 2009) and Turkey (Nebioglu et al., 2012). However, the validation and application of the full and BSCS scales in China is still in its infancy. An initial study conducted in Chinese amongst college students in Wuhan suggested that the full version of SCS supports a five-factor construct scale (Tan and Guo, 2008), which was then used to examine the patterns of mobile phone usage amongst the students (Jiang and Zhao, 2016). Unger et al. (2016) proposed validating Tangney and associates' SCS in mainland China, and attempted to investigate the psychometric properties of SCS and BSCS using 371 Chinese college students between 17 and 23 years old. They found that both scales had a satisfactory internal consistency and a reasonable goodness of fit for the five-factor construct. They concluded that the BSCS was preferable to the SCS as it had a strong correlation with the full scale but saved time and had a higher rate of return.

The aim of this study is to re-examine the 13-item BSCS in two ways. First, it evaluates the issue of dimensionality of the BSCS. The literature continues to be controversial with regard to the multi-factor structure of the BSCS. The original scale developers and the subsequent validation studies replicated the five-factor structure, i.e., general capacity for self-discipline (5 items), inclination toward deliberate or non-impulsive action (3 items), healthy habits (2 items), self-regulation in service to build a strong work ethic (2 items), and reliability (1 item) (Tangney et al., 2004; Unger et al., 2016). Since the introduction of SCSs in 2004, scholars have offered other conceptualizations of self-control with different dimensions (Fulford et al., 2008; Friese and Hofmann, 2009) and have proposed different conceptualizations of two-factor structures on the basis of the existing 13-item BSCS, such as general self-discipline (9 items) and impulse control (4 items) (Ferrari et al., 2009). Maloney et al. (2012) proposed an 8-item BSCS, focusing on impulsivity (4 items) and restrain (4 item). Alternatively, a 10-item BSCS, emphasizing inhibition (6 items) and initiation (4 items) was suggested by de Ridder et al. (2011). Lindner et al. (2015) attempted to evaluate the above two-dimensional BSCS specifications, but could not demonstrate which conceptualization of the BSCS was more appealing. Hence, evaluating the dimensionality of the Chinese version of BSCS warrants attention.

Second, the Chinese version of BSCS's psychometric properties is subject to further investigation. Unger et al. (2016) have attempted to validate the Chinese version of BSCS in China, however, their study with potential limitations like small sample size and inadequate evaluation of criterion validity. Hence, the design of this study in particular, pays closer attention to the criterion validity of self-control with other construct-related scales related to the conceptualization of self-control. Furthermore, some low factors loadings of the scale items needed retesting to confirm whether they need replaced.

MATERIALS AND METHODS

Participants

This cross-sectional study recruited 903 respondents from Huashang College, Guangdong University of Business Studies, located in the southern part of China. The gender ratio of the sample (792 females to 111 males) matched that of the official school record, i.e., over 80% of the students enrolled in the university were female. The average age of the respondents was 20.56 years (SD = 2.753). Student sample profiles of this study matched those of the original scale developers who had recruited 28% male and 72% female and 19% male and 81% female university students in study 1 and study 2 samples, respectively (Tangney et al., 2004).

Measures

The full version of the SCS comprises of 36 items. The original scale developers proposed using the shortened version, the BSCS, which contains 13 items, including 1, 2, 3, 4, 6, 13, 17, 22, 28, 29, 30, 31, and 32. These 13 items were rated on a 5-point Likert scale ranging from 1, *not at all like me*, to 5, *very much like me*. Eight items, including 2, 3, 4, 6, 17, 28, 29, and 31 had reversed scores (Tangney et al., 2004). The reversed items were re-coded in the dataset prior to the analysis.

The Chinese version of the BSCS was adapted from Unger et al. (2016). We recruited two translators who were fluent in both English and Chinese to cross-check the translated versions to verify whether the original English and Chinese versions were identical (Brislin, 1970). To further ensure that the translated versions were free from any cultural biases, two pilot studies were conducted in Xi'an and Guangzhou, located in northern China and southern China, respectively. Each pilot study involved five mainland Chinese university students from diverse academic backgrounds, ranging from accountancy and management to sports sciences, computer sciences, and the social sciences. None of the participants reported any difficulties in understanding and answering the questions. Data from the pilot studies were excluded in the dataset.

Procedures

The research team used the announcement function in the school-based intranet smartphone application available in both iOS and Android operating systems to recruit students voluntarily participate in an online self-reported survey related to self-control, well-being and Internet usage from June to July 2018. On the questionnaire page, students were fully informed the background of the study and we obtained informed consent from the participants prior to allow them to complete the self-administered questionnaire. The respondents were only able to submit the completed questionnaire once. Each participant spent around 10 min completing the questionnaire. The data that we collected were anonymous. The study was approved by the ethical committee of the Huashang College, Guangdong University of Business Studies. The entire research process and data collection procedure also complied with the ethical standards of the Declaration of Helsinki and the relevant government policies stipulated in the Article 14 of Chapter III, Statistics Law of the People's Republic of China.

Various psychometric testing tools and validated instruments were used to examine the BSCS. The internal consistency of the BSCS was assessed by Cronbach's alpha (Cronbach, 1951), McDonald's Omega (McDonald, 1999; Zinbarg et al., 2005; Revelle and Zinbarg, 2009) and the corrected item-total correlations between all the 13 items were examined (Hair, 2010; Tabachnick, 2013). The criterion validity was evaluated with other validation constructs or measurements reported in relevant studies on self-control as well as the item-to-scale correlations (Beaton et al., 2000; Loewenthal, 2001). According to Tangney et al. (2004) and Unger et al. (2016), the SCS is positively correlated with self-esteem, happiness, quality of life, and well-being, but has significant moderate negative correlations with psychometric instruments related to psychological problems and symptoms of psychopathology, such as the 12-item General Health Questionnaire (GHQ-12). Owing to the availability of the validated Chinese scales and the length of the questionnaires, five well-established instruments were used to evaluate the criterion validity of the BSCS: The GHQ-12 evaluated by twelve items (with five reversed items) to assess the severity of health related problems using a 4-point Likert-type scale. Respondents with high scores indicate worse health (Goldberg and Williams, 1991); Rosenberg Self-esteem Scale (RSES) consists of ten statements (with five reversed items) evaluated by 4-point Likert-type scale, with 1 = *strongly disagree* and 4 = *strongly agree*. High scores refer to high level of self-esteem (Rosenberg, 1965; Rosenberg et al., 1989); Satisfaction with Life Scale (SWLS) comprised of five items with 7-point Likert-type scale (1 = *strongly disagree*; 7 = *strongly agree*). High scores signify the respondents highly satisfied with their life (Diener et al., 1985; Pavot et al., 1991; Pavot and Diener, 1993, 2008); Subjective Happiness Scale (SHS) consists of four statements measured by 7-point Likert-type scale. High scores mean happier (Lyubomirsky and Lepper, 1999); and WHO (Five) Well-Being Index (WHO-5) comprised of five items with 6-point Likert-type scale (0 = *at no time*; 5 = *all of the time*), high score indicates high level of well-being (Bech et al., 2003; Bech, 2004, 2012). In addition to the original 13-item BSCS (Tangney et al., 2004) and several basic demographic questions, the participants were asked to complete a questionnaire with 51 items.

The evaluation of the scale's factorial validity was based on exploratory factor analysis (EFA). There are controversies about the rotation method used in the EFA (Jennrich and Sampson, 1966). Current BSCS studies use different EFA extraction methods, thereby giving rise to controversies with regard to the multi-factor structure. For example, a recent study used principal components with direct oblimin rotation (Maloney et al., 2012); Ferrari et al. (2009) used the maximum likelihood process with varimax rotation. However, the original scale developers used principal components with varimax to trim the SCS scale from 36 to 13 items (Tangney et al., 2004). The varimax is a commonly used orthogonal factor rotation method for simplified factor structures (Hair, 2010). Hence, we adopted principal components with varimax as an EFA rotation method, which is the same as the originally developed scale, to evaluate the Chinese version of the BSCS. Due to a relatively large sample size, i.e., over 350 respondents in this study; hence, an item with a factor loading over 0.50 can be interpreted as having practical significance (Hair, 2010).

Confirmatory factor analysis (CFA) was used to examine the construct validity of the scale (Jöreskog, 1969; Loewenthal, 2001; Brown, 2014). Although it has been argued that the maximum likelihood estimator is inappropriate for the ordinal nature of the BSCS (Lionetti et al., 2016), existing studies have predominantly used it in CFA (de Ridder et al., 2011; Maloney et al., 2012; Lindner et al., 2015; Unger et al., 2016). To address this issue, CFA has been conducted to examine the factor structure of the BSCS using the diagonally weighted least squares (DWLS) method. The usage of the DWLS estimator, which is suitable for ordinal items constructed scales, and is an effective tool for evaluating the dimensionality and psychometric properties of BSCS in the following two reasons. The BSCS as a latent construct is estimated by Likert scale items consisting of ordinal data, and the DWLS method is regarded as having a less biased and more optimal fit (DiStefano and Morgan, 2014; Li, 2016; Lionetti et al., 2016). In addition, the results of this study can be directly compared with other BSCS validation studies using frequentist estimations (de Ridder et al., 2011; Maloney et al., 2012; Nebioglu et al., 2012; Lindner et al., 2015; Unger et al., 2016). The model fit and cut-off criteria were evaluated on the basis of the following cut-off values; a comparative fit index (CFI) and a Tucker-Lewis fit index (TLI) of over 0.950, a standardized root mean square residual (SRMR) under 0.08 and an root mean square error of approximation (RMSEA) under 0.06, which were considered good fits (Browne and Cudeck, 1993; Hu and Bentler, 1999; Schreiber et al., 2006; Hair, 2010; Bass et al., 2016). An acceptable model can also be indicated by $\chi^{2/}\text{df} \leq 3$ due to the large sample size (Bentler and Bonett, 1980; Kline, 2005). The analyses were implemented with the IBM SPSS 25.0 and the lavaan package version 0.6-3 (Rosseel, 2012) in R version 3.5.2.

RESULTS

Internal Consistency

Table 1 shows the means, standard deviations, skewness, kurtosis, corrected item-total correlations, and Cronbach's alpha if items were deleted of the BSCS (*N* = 903). The mean score for the BSCS among all the respondents, male and female were

TABLE 1 | Descriptive statistics for 13-item BSCS scale.

Item	Mean	*SD*	Skewness	Kurtosis	Corrected item-total correlation	Cronbach's alpha if items were deleted
BSCS1	3.38	0.970	−0.271	−0.077	0.440	0.782
BSCS2	2.79	1.018	0.044	−0.408	0.534	0.773
BSCS3	2.58	1.089	0.250	−0.454	0.550	0.771
BSCS4	3.20	1.063	−0.161	−0.488	0.473	0.779
BSCS6	3.67	1.150	−0.501	−0.614	0.459	0.780
BSCS13	3.78	1.052	−0.601	−0.260	0.196	0.803
BSCS17	1.82	0.873	0.830	0.176	0.077	0.808
BSCS22	3.04	1.058	−0.076	−0.425	0.432	0.782
BSCS28	2.44	1.060	0.402	−0.379	0.386	0.786
BSCS29	2.87	1.038	0.018	−0.375	0.552	0.772
BSCS30	3.20	0.982	−0.091	−0.205	0.398	0.785
BSCS31	2.82	1.101	0.169	−0.590	0.515	0.774
BSCS32	3.19	1.113	−0.059	−0.648	0.500	0.776

38.77 (SD = 7.32), 39.33 (SD = 7.35), and 38.68 (SD = 7.32), respectively, which is similar to that reported in the original study (Tangney et al., 2004). No significant differences and relationship were observed in the scale scores on sex of the respondent based on the independent-sample *t*-test and correlation results. The corrected item-to-total correlations in the 13-item BSCS ranged from 0.077 to 0.550. The following two items reported values lower than 0.300: BSCS17 (0.077) and BSCS13 (0.196). This finding was addressed in the subsequent EFA while evaluating the scale's factorial validity. The Cronbach's alpha of the BSCS in this study was 0.80, replicating the original BSCS Cronbach's alpha values, i.e., 0.83 and 0.85 in studies 1 and 2, respectively (Tangney et al., 2004). The results suggested that the scale is highly reliable in terms of internal consistency.

Criterion Validity

According to Tangney et al. (2004), self-control is one of the most powerful and beneficial aspects of the human psyche, and is positively related to happiness and health. The BSCS is demonstrated to have significant moderate positive correlations with self-esteem, quality of life and well-being (Tangney et al., 2004; Unger et al., 2016). As shown in **Table 2**, the Chinese version of the BSCS also showed significant moderate correlations with RSES ($r = 0.459$, $p < 0.001$), SWLS ($r = 0.302$, $p < 0.001$), SHS ($r = 0.332$, $p < 0.001$), and WHO-5 ($r = 0.243$, $p < 0.001$).

TABLE 2 | Correlation between 13-item BSCS scale in relation to other validation constructs.

Other construct-related scales	BSCS
Rosenberg Self-esteem Scale (RSES)	0.459***
Satisfaction with Life Scale (SWLS)	0.302***
Subjective Happiness Scale (SHS)	0.332***
WHO (Five) Well-Being Index (WHO-5)	0.243***
12-item General Health Questionnaire (GHQ-12)	−0.422***

****p < 0.001.*

To further evaluate the criterion validity of the BSCS, whether the scale demonstrated a negative relationship with the psychological symptoms-related scale was also assessed. The results of the correlation show that the Chinese version of the BSCS demonstrated a significant moderate negative relationship with GHQ-12 ($r = −0.422$, $p < 0.001$). This finding also replicated the existing studies' findings in terms of the direction and magnitude of the scales related to mental disorder (Tangney et al., 2004; Unger et al., 2016). **Table 3** shows the correlations between specific items and other construct-related scales. However, BSCS17 in particular, showed a very weak association with other scales, suggesting an opposite correlation orientation in the RSES, SHS, and GHQ-12 scales. In short, the 13-item BSCS demonstrated good criterion validity with the other validation constructs.

Factorial Validity

Table 4 shows the results of the EFA using principal component analysis with varimax rotation as adopted by the original scale

TABLE 3 | Correlations between the BSCS items and other construct-related scales.

Item	BSCS	RSES	SWLS	SHS	WHO-5	GHQ-12
BSCS1	0.543***	0.285***	0.301***	0.247***	0.274***	−0.302***
BSCS2	0.629***	0.258***	0.129***	0.145***	0.124***	−0.210***
BSCS3	0.649***	0.284**	0.216***	0.192***	0.154***	−0.232***
BSCS4	0.582***	0.328***	0.114***	0.231***	0.060	−0.279***
BSCS6	0.579***	0.260***	0.028	0.145***	0.033	−0.267***
BSCS13	0.332***	0.192***	0.113***	0.162***	0.065*	−0.200***
BSCS17	0.195***	−0.085*	−0.009	−0.075*	0.021	0.086**
BSCS22	0.546***	0.230***	0.274***	0.192***	0.219***	−0.210***
BSCS28	0.506***	0.176***	0.100**	0.125***	0.106**	−0.179***
BSCS29	0.646***	0.347***	0.243***	0.247***	0.185***	−0.342***
BSCS30	0.508***	0.310***	0.332***	0.230***	0.244***	−0.273***
BSCS31	0.622***	0.264***	0.172***	0.197***	0.141***	−0.217***
BSCS32	0.610***	0.318***	0.111***	0.252***	0.091**	−0.282***

**p < 0.05, **p < 0.01, ***p < 0.001.*

TABLE 4 | Factor loading for the Brief Self-Control Scale.

Item	13-item BSCS with 5-factor structure					11-item BSCS with 4-factor structure			
	F1	F2	F3	F4	F5	F1	F2	F3	F4
BSCS1	0.261	0.754	−0.050	0.091	0.117	0.278	0.752	−0.062	0.104
BSCS2	0.242	0.169	0.733	0.196	0.008	0.197	0.175	0.757	0.199
BSCS3	0.222	0.262	0.726	0.156	0.137	0.176	0.271	0.731	0.200
BSCS4	0.644	−0.023	0.399	0.039	0.080	0.594	−0.044	0.466	0.052
BSCS6	0.724	0.055	0.248	0.004	−0.166	0.716	0.046	0.300	−0.070
BSCS13	0.470	0.380	−0.259	−0.108	−0.422	–	–	–	–
BSCS17	−0.012	0.052	0.063	0.101	0.909	–	–	–	–
BSCS22	0.018	0.784	0.307	0.014	0.044	0.013	0.788	0.293	0.008
BSCS28	0.114	0.045	0.101	0.813	0.186	0.069	0.017	0.177	0.874
BSCS29	0.253	0.203	0.235	0.724	−0.049	0.312	0.246	0.200	0.655
BSCS30	−0.083	0.726	0.262	0.220	−0.163	−0.026	0.768	0.201	0.129
BSCS31	0.671	0.128	0.027	0.322	0.198	0.693	0.137	0.026	0.342
BSCS32	0.675	0.070	0.117	0.300	−0.086	0.762	0.122	0.060	0.199

developers, who extracted five factors from the scale (Tangney et al., 2004). The explanation power of the factors relative to the total variance is explained as follows: Factor 1 explaining 17.9% of the variance consists of five items, including BSCS4, BSCS6, BSCS13, BSCS31, and BSCS32, related to the general capacity for self-discipline. BSCS13 has a factor loading of 0.470 only, which is slightly lower than the practical and significant value of 0.500; Factor 2, which is related to inclination toward deliberate/non-impulsive action consists of items BSCS1, BSCS22, and BSCS30, yielding 15.6% explanation power; Factor 3 explaining 12.3% of the variance, which is related to healthy habits consists of BSCS2 and BSCS3; Factor 4, which is related to self-regulation in service for building a strong work ethic consists of items BSCS28 and BSCS29, with 11.7% explanation power; and Factor 5 is related to reliability with item BSCS17 explaining 9.0% of the variance. The above results are identical to those of the five-factor model suggested in the original study (Tangney et al., 2004). By removing BSCS13 and BSCS17 from the scale, the EFA results of the 11-item BSCS with a four-factor structure suggested that all of the factor loadings in each factor ranged from 0.594 to 0.974 and that it supported a scale construction. The EFA results showed that the assertion of a two-factor structure suggested in the BSCS literature (Ferrari et al., 2009; de Ridder et al., 2011; Maloney et al., 2012) is not supported in this study.

Construct Validity

Table 5 shows the results of the CFA of the BSCS. Model 1 evaluated all of the 13-items of BSCS based on a single factor. The results indicated that the scale did not fit the model well, with χ^2 (1362.277) = 65, $p < 0.001$, SRMR = 0.106, CFI = 0.873, TLI = 0.847, and RMSEA = 0.149. The five-factor model suggested in the original scale (Tangney et al., 2004) failed to obtain any results, as the fifth factor only consisted of one item, and hence the model was not identified. Model 2, which was based on the suggestions of Ferrari et al. (2009), reconceptualized the BSCS into a two-factor structure, which included general self-discipline (BSCS2, BSCS3, BSCS4, BSCS6, BSCS13, BSCS17, BSCS29, and BSCS30) and impulse control (BSCS1, BSCS28, BSCS31, and BSCS32). The CFA results also reported a poor model fit, with χ^2 (1356.189) = 64, $p < 0.001$, SRMR = 0.106, CFI = 0.873, TLI = 0.845, and RMSEA = 0.150. Likewise, the results in Model 3 also demonstrated the other 10-item, two-factor structure of the BSCS proposed by de Ridder et al. (2011), namely, inhibition (BSCS1, BSCS2, BSCS6, BSCS17, BSCS29, and BSCS31) and initiation (BSCS3, BSCS22, BSCS28, and BSCS30). However, it failed to fulfill the cut-off criteria for a good model fit, as χ^2 (638.066) = 34, $p < 0.001$, SRMR = 0.093, CFI = 0.904, TLI = 0.873, and RMSEA = 0.140. Model 4 evaluated a recent study that suggested an 8-item BSCS with a two-factor structure, namely, restraint (BSCS1, BSCS2, BSCS17, and BSCS22) and impulsivity (BSCS6, BSCS28, BSCS31, and BSCS32) derived from samples used in the Midwestern United States (Maloney et al., 2012). The results indicated that the two-factor structure also failed to fulfill the criteria for goodness of fit, with χ^2 (346.287) = 19, $p < 0.001$, SRMR = 0.092, CFI = 0.886, TLI = 0.831, and RMSEA = 0.138.

TABLE 5 | Confirmatory factor analysis of the BSCS.

Model	No. of factors	χ^2	df	RMSEA	CFI	TLI	SRMR
BSCS (13 items)							
1	1	1362.277***	65	0.149	0.873	0.847	0.106
2	2	1356.189***	64	0.150	0.873	0.845	0.106
BSCS (10 items)							
3	2	638.066***	34	0.140	0.904	0.873	0.093
BSCS (8 items)							
4	2	346.287***	19	0.138	0.886	0.831	0.092
BSCS (11 items)							
5	4	125.391***	38	0.050	0.991	0.986	0.039
6[a]	4	106.626***	37	0.046	0.992	0.989	0.036

[a] *Includes the covariance between the error terms for items BSCS4 and BSCS31.*
****p < 0.001.*

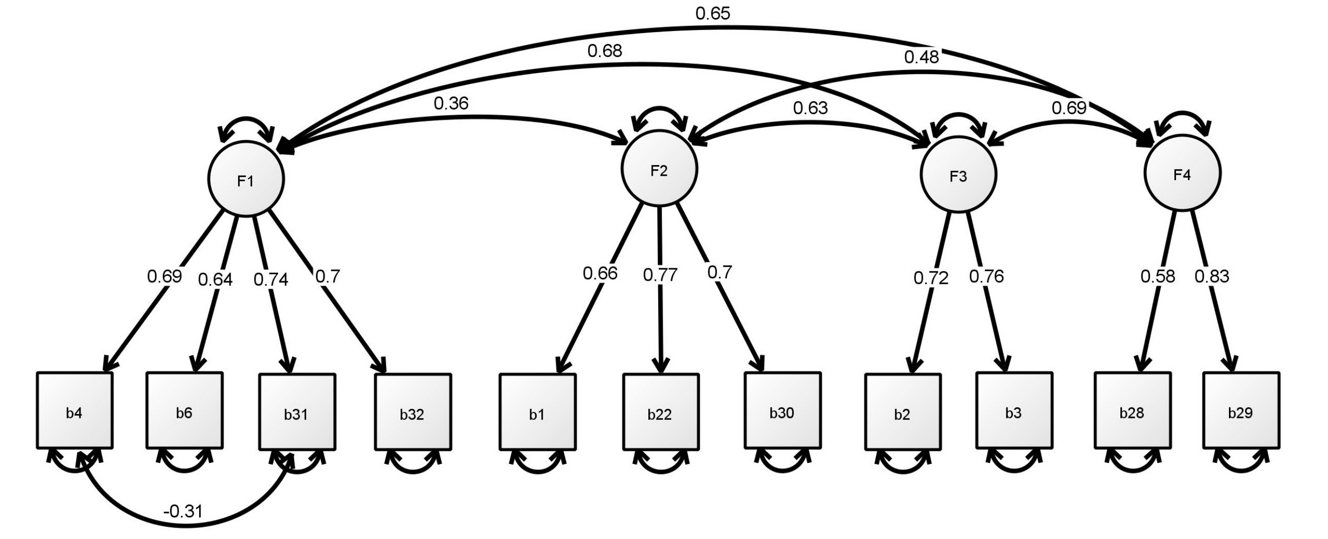


FIGURE 1 | Final standardized model of the 11-item BSCS. F1, self-discipline; F2, impulsivity; F3, healthy habits; F4, self-regulation.

We propose a shortened version of the BSCS by removing two items, namely, BSCS13, factor 1 related to general capacity for self-discipline, and BSCS17, factor 5 related to reliability, based on the findings of prior analyses. The 11-item BSCS consisted of a four-factor structure, namely, F1) self-discipline: BSCS4, BSCS6, BSCS31, and BSCS32; F2) impulsivity: BSCS1, BSCS22, and BSCS30; F3) healthy habits: BSCS2 and BSCS3; and F4) self-regulation: BSCS28 and BSCS29. The CFA in Model 5 was conducted without correlating the error terms and the results were very close to the criteria of a goodness of fit other than χ^2/df value = 3.30. Model 6 re-evaluated the 11-item BSCS, with the error correlations based on the modification indices, and it included one covariance factor between the error terms for BSCS4 and BSCS31. The data suggest that the shortened version is suitable for a four-factor scale with *post hoc* modification. The results indicated good model fit, as χ^2 (106.626)/37 = 2.88, SRMR = 0.036, CFI = 0.992, TLI = 0.989, RMSEA = 0.046. In addition, the omega total (ωt) recorded 0.86, which indicated above the acceptable range. **Figure 1** presents the final standardized model 1. In short, the results suggest that the 11-item BSCS comprised of items 1, 2, 3, 4, 6, 22, 28, 29, 30, 31, and 32 with a four-factor structure is an appropriate measure of self-control amongst the Chinese university student population.

DISCUSSION

The main contribution of this study is the re-examination of the psychometric properties and dimensionality of the BSCS in mainland China. The findings of this study suggest that a shortened version of the 11-item BSCS with a four-factor structure had better psychometric properties and good model fit in the CFA of Chinese college students. The revised version removed BSCS13 and BSCS17, and included the following four factors: self-discipline (BSCS4, BSCS6, BSCS31, and BSCS32), impulsivity (BSCS1, BSCS22, and BSCS30), healthy habits (BSCS2 and BSCS3) and self-regulation (BSCS28 and BSCS29). In terms of psychometric properties, the revised Chinese translated version of the 11-item BSCS had a high degree of internal consistency with a Cronbach's alpha of 0.81. Both the 11-item BSCS and the 13-item BSCS demonstrated very strong and significant positive correlations with $r = 0.988$, $p < 0.001$. The revised scale also had good criterion validity with other well-established scales that are theoretically and conceptually related to self-control. The 11-item BSCS displayed good criterion validity with other construct-related scales and showed a significant moderate relation with self-esteem (RSES, $r = 0.469$), quality of life (SWLS, $r = 0.305$; WHO-5, $r = 0.246$), happiness (SHS, $r = 0.337$), and minor psychological disorders (GHQ-12, $r = -0.428$).

With regard to the controversy related to the dimensionality of BSCS, we had examined the five-factor (Tangney et al., 2004; Unger et al., 2016), two-factor (Ferrari et al., 2009; de Ridder et al., 2011; Maloney et al., 2012) and single factor constructs (Lindner et al., 2015) using CFA. The five-factor constructs of the BSCS suggested in the original scale failed to yield CFA results as the fifth factor was potentially problematic as it consisted of only one item. The findings show that the single and two-factor constructs presented in Models 1, 2, 3, and 4 failed to achieve the adequate model fit criteria. The four-factor constructs without correlating the error terms in Model 5 with RMSEA, CFI, TLI, and SRMR values were a good fit model, but χ^2 was significant ($p < 0.001$) probably due to the effects of the large sample size (Bentler and Bonett, 1980; Kline, 2005); hence, after the covariance in the error terms based on modification indices (Shah and Goldstein, 2006; Cole et al., 2008), Model 6 was good model fit for the constructs of the BSCS (**Appendix**). In short, the proposed scale in this study in general retained the original factors proposed by the original scale developers (Tangney et al., 2004). It avoided the problem of artificially rearrange the factor structure without based on any theoretical justifications.

There are several potential limitations associated with this study. First, only limited number of self-control-related scales to verify the criterion validity of the BSCS in this study. Tangney et al. (2004) used measures such as the Marlowe–Crowne Social Desirability scale, the Eating Disorder Inventory, the Michigan Alcohol Screening Test, and the Symptom Checklist 90 to evaluate the BSCS. Owing to the availability of reliable Chinese translated scales and the length of the questionnaire, we adopted other well established construct-related scales, such as the RSES, SWLS, SHS, WHO-5, and GHQ12 that are commonly used or discussed in BSCS validation studies and the literature on self-control (Rothbaum et al., 1982; Tangney, 1991, 1995; Baumeister, 1994; Tangney et al., 1996, 2004; Eisenberg et al., 1998; Fabes et al., 1999; Unger et al., 2016). The findings of this study consistently demonstrate that the BSCS possesses good criterion validity in terms of magnitude and direction with other self-control related scales suggested in the literature.

Second, the sample used in this study may also limit the generalizability of the findings given that the respondents were recruited from one Chinese university with large proportion of female population. However, this limitation may have been compensated by a relatively large sample size in the university setting with reference to the other BSCS related studies. As such, Tangney et al. (2004) managed to recruit only 351 and 255 students in their studies to develop the BSCS. More importantly, we have computed additional confirmative factor analysis on both male and female participants with the 11-item BSCS. The analysis indicated the same results as we presented in Model 6, as male students with χ^2 (37.845)/37 = 1.02, SRMR = 0.058, CFI = 0.999, TLI = 0.999, and RMSEA = 0.014 (n = 111), while female students with χ^2 (111.366)/37 = 3.0, SRMR = 0.039, CFI = 0.991, TLI = 0.987, and RMSEA = 0.050 (n = 792). Both results fulfilled all the cut-off criteria for good model fit.

FUTURE RESEARCH

To evaluate the construct validity of the scale, further studies should examine and verify the four dimensional 11-item BSCS in other Chinese populations and focus on further confirming BSCS's validity with regard to the general public and other populations. Future studies need to make use of other population samples to establish the BSCS's wider applicability in the future. Besides, schools, reformative agencies, and practitioners could use the BSCS along with intervention programmes to evaluate its effectiveness in strengthening participants' self-control in the Chinese context. Finally, the concept of self-control is essential in the social and psychological context. It is conceptually related to many theories and applications, such as criminology, positive psychology, subjective well-being, and quality of life. Further exploration may provide further insights into accurately describing human behavior.

CONCLUSION

To conclude, the findings show that the BSCS is reliable in Chinese culture and is applicable to Chinese college populations. The results suggested that an 11-item BSCS (without BSCS13 and BSCS17) with a four-factor structure fulfilled all the cut-off criteria for good model fit in CFA. A validated Chinese version of the BSCS provides a comprehensive and handy measure for broader research in the context of mainland China or the Chinese diaspora.

DATA AVAILABILITY STATEMENT

The dataset used and/or generated for this study is available from the corresponding author on reasonable request.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of Statistics Law of the People's Republic of China. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethical Committee of the Huashang College, Guangdong University of Business Studies.

AUTHOR CONTRIBUTIONS

SF: study design, data collection, data analysis, data interpretation, and manuscript preparation. CK: study design and manuscript preparation. QH: study design and data collection.

REFERENCES

Ainslie, G. (1975). Specious reward: a behavioral theory of impulsiveness and impulse control. *Psychol. Bull.* 82, 463–496. doi: 10.1037/h0076860

Bass, M., Dawkin, M., Muncer, S., Vigurs, S., and Bostock, J. (2016). Validation of warwick-edinburgh mental well-being scale (WEMWBS) in a population of people using Secondary Care Mental Health Services. *J. Mental Health* 25, 323–329. doi: 10.3109/09638237.2015.1124401

Baumeister, R. F. (1994). *Losing Control: How and Why People Fail at Self-Regulation*. San Diego, CA: Academic Press.

Baumeister, R. F. (2016). "Self-Control and Ego Depletion," in *Handbook of Self-Regulation : Research, Theory, and Applications*, 3rd Edn, eds K. D. Vohs, and R. F. Baumeister (New York, NY: The Guilford Press), 42–61.

Beaton, D. E., Bombardier, C., Guillemin, F., and Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 25, 3186–3191. doi: 10.1097/00007632-200012150-00014

Bech, P. (2004). Measuring the dimensions of psychological general well-being by the WHO-5. *QoL Newslett.* 32, 15–16.

Bech, P. (2012). *Clinical Psychometrics*. Chichester: Wiley-Blackwell.

Bech, P., Olsen, L. R., Kjoller, M., and Rasmussen, N. K. (2003). Measuring well-being rather than the absence of distress symptoms: a comparison of the SF-36

Mental Health subscale and the WHO-Five well-being scale. *Int. J. Methods Psychiatr. Res.* 12, 85–91. doi: 10.1002/mpr.145

Bentler, P. M., and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88, 588–606. doi: 10.1037/0033-2909.88.3.588

Bertrams, A., and Dickhäuser, O. (2009). Messung dispositioneller Selbstkontroll-Kapazität. *Diagnostica* 55, 2–10. doi: 10.1026/0012-1924.55.1.2

Brevers, D., Foucart, J., Verbanck, P., and Turel, O. (2017). Examination of the validity and reliability of the French version of the Brief Self-Control Scale. *Can. J. Behav. Sci.* 49, 243–250. doi: 10.1037/cbs0000086

Brislin, R. W. (1970). Back-translation for cross-cultural research. *J. Cross Cult. Psychol.* 1, 185–216. doi: 10.1177/135910457000100301

Brown, T. A. (2014). *Confirmatory Factor Analysis for Applied Research*, 2nd Edn. New York, NY: Guilford Publications.

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen, and J. S. Long, (Newburyk Park: Sage), 136–162.

Cole, D. A., Ciesla, J., and Steiger, J. (2008). The insidious effects of failing to include design-driven residuals in latent-variable covariance structure analysis. *Psychol. Methods* 12, 381–398. doi: 10.1037/1082-989X.12.4.381

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/bf02310555

de Ridder, D. T. D., de Boer, B. J., Lugtig, P., Bakker, A. B., and van Hooft, E. A. J. (2011). Not doing bad things is not equivalent to doing the right thing: distinguishing between inhibitory and initiatory self-control. *Pers. Individ. Differ.* 50, 1006–1011. doi: 10.1016/j.paid.2011.01.015

de Ridder, D. T. D., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., and Baumeister, R. F. (2012). taking stock of self-control: a meta-analysis of how trait self-control relates to a wide range of behaviors. *Pers. Soc. Psychol. Rev.* 16, 76–99. doi: 10.1177/1088868311418749

Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess.* 49, 71–75. doi: 10.1207/s15327752jpa4901_13

DiStefano, C., and Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Struc. Equa. Mod.* 21, 425–438. doi: 10.1080/10705511.2014.915373

Duckworth, A. L., and Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *J. Res. Pers.* 45, 259–268. doi: 10.1016/j.jrp.2011.02.004

Eisenberg, N., Fabes, R. A., Shepard, S. A., Murphy, B. C., Jones, S., and Guthrie, I. K. (1998). Contemporaneous and longitudinal prediction of children's sympathy from dispositional regulation and emotionality. *Dev. Psychol.* 34, 910–924. doi: 10.1037/0012-1649.34.5.910

Fabes, R. A., Eisenberg, N., Jones, S., Smith, M., Guthrie, I., Poulin, R., et al. (1999). Regulation, emotionality, and preschoolers'. Socially competent peer interactions. *Child Dev.* 70, 432–442. doi: 10.1111/1467-8624.00031

Ferrari, J. R., Stevens, E. B., and Jason, L. A. (2009). The relationship of self-control and abstinence maintenance: an exploratory analysis of self-regulation. *J. Groups Addict. Recov.* 4, 32–41. doi: 10.1080/15560350802712371

Friese, M., and Hofmann, W. (2009). Control me or I will control you: impulses, trait self-control, and the guidance of behavior. *J. Res. Pers.* 43, 795–805. doi: 10.1016/j.jrp.2009.07.004

Fulford, D., Johnson, S. L., and Carver, C. S. (2008). Commonalities and differences in characteristics of persons at risk for narcissism and mania. *J. Res. Personality* 42, 1427–1438. doi: 10.1016/j.jrp.2008.06.002

Goldberg, D. P., and Williams, P. (1991). *A User's Guide to the General Health Questionnaire*. Berkshire: NFER-NELSON.

Gottfredson, M. R. (1990). *A General Theory of Crime*. Stanford, CA: Stanford University Press.

Grasmick, H. G., Tittle, C. R., Bursik, R. J., and Arneklev, B. J. (1993). Testing the core empirical implications of gottfredson and Hirschi's general theory of crime. *J. Res. Crime Delinquen.* 30, 5–29. doi: 10.1177/0022427893030001002

Hair, J. F. (2010). *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struc. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Jennrich, R. I., and Sampson, P. F. (1966). Rotation for simple loadings. *Psychometrika* 31, 313–323. doi: 10.1007/bf02289465

Jiang, Z., and Zhao, X. (2016). Self-control and problematic mobile phone use in Chinese college students: the mediating role of mobile phone use patterns. *BMC Psychiatry* 16:416. doi: 10.1186/s12888-016-1131-z

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34, 183–202. doi: 10.1007/bf02289343

Kendall, P. C., and Wilcox, L. E. (1979). Self-control in children: development of a rating scale. *J. Consult. Clin. Psychol.* 47, 1020–1029. doi: 10.1037/0022-006X.47.6.1020

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford Press.

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behav. Res. Methods* 48, 936–949. doi: 10.3758/s13428-015-0619-7

Lindner, C., Nagy, G., and Retelsdorf, J. (2015). The dimensionality of the Brief Self-Control Scale—An evaluation of unidimensional and multidimensional applications. *Pers. Individ. Differ.* 86, 465–473. doi: 10.1016/j.paid.2015.07.006

Lionetti, F., Keijsers, L., Dellagiulia, A., and Pastore, M. (2016). Evidence of factorial validity of parental knowledge, control and solicitation, and adolescent disclosure scales: when the ordered nature of Likert scales matters. *Front. Psychol.* 7:941. doi: 10.3389/fpsyg.2016.00941

Loewenthal, K. M. (2001). *An Introduction to Psychological Tests and Scales*. Philadelphia, PA: Psychology Press.

Lyubomirsky, S., and Lepper, H. S. (1999). A measure of subjective happiness: preliminary reliability and construct validation. *Soc. Indicat. Res.* 46, 137–155. doi: 10.1023/a:1006824100041

Maloney, P. W., Grawitch, M. J., and Barber, L. K. (2012). The multi-factor structure of the Brief Self-Control Scale: discriminant validity of restraint and impulsivity. *J. Res. Pers.* 46, 111–115. doi: 10.1016/j.jrp.2011.10.001

Malouf, E. T., Schaefer, K. E., Witt, E. A., Moore, K. E., Stuewig, J., and Tangney, J. P. (2014). The brief self-control scale predicts Jail Inmates' recidivism, substance dependence, and post-release adjustment. *Pers. Soc. Psychol. Bull.* 40, 334–347. doi: 10.1177/0146167213511666

McDonald, R. P. (1999). *Test Theory: a Unified Treatment*. London: L. Erlbaum Associates.

Metcalfe, J., and Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychol. Rev.* 106, 3–19. doi: 10.1037/0033-295X.106.1.3

Mischel, H. N., and Mischel, W. (1983). The development of children's knowledge of self-control strategies. *Child Dev.* 54, 603–619. doi: 10.2307/1130047

Mischel, W. (1974). "Processes in delay of gratification," in *Advances in Experimental Social Psychology*, ed. L. Berkowitz (Cambridge, MA: Academic Press), 249–292. doi: 10.1016/s0065-2601(08)60039-8

Nebioglu, M., Konuk, N., Akbaba, S., and Eroglu, Y. (2012). The investigation of validity and reliability of the turkish version of the brief self-control scale. *Klinik Psikofarmakoloji Bülteni* 22, 340–351. doi: 10.5455/bcp.20120911042732

Pavot, W., and Diener, E. (1993). Review of the satisfaction with life scale. *Psychol. Assess.* 5, 164–172. doi: 10.1037/1040-3590.5.2.164

Pavot, W., and Diener, E. (2008). The satisfaction with life scale and the emerging construct of life satisfaction. *J. Posit. Psychol.* 3, 137–152. doi: 10.1080/17439760701756946

Pavot, W., Diener, E., Colvin, C. R., and Sandvik, E. (1991). Further validation of the Satisfaction with Life Scale: evidence for the cross-method convergence of well-being measures. *J. Pers. Assess.* 57, 149–161. doi: 10.1207/s15327752jpa5701_17

Piquero, A. R., and Bouffard, J. A. (2007). Something old, something new: a preliminary investigation of Hirschi's redefined self-control. *Justice Q.* 24, 1–27. doi: 10.1080/07418820701200935

Revelle, W., and Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: comments on Sijtsma. *Psychometrika* 74, 145–154. doi: 10.1007/s11336-008-9102-z

Rosenberg, M. (1965). *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press.

Rosenberg, M., Schooler, C., and Schoenbach, C. (1989). Self-Esteem and adolescent problems: modeling reciprocal effects. *Am. Sociol. Rev.* 54, 1004–1018. doi: 10.2307/2095720

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48:36. doi: 10.18637/jss.v048.i02

Rothbaum, F., Weisz, J. R., and Snyder, S. S. (1982). Changing the world and changing the self: a two-process model of perceived control. *J. Pers. Soc. Psychol.* 42, 5–37. doi: 10.1037/0022-3514.42.1.5

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338. doi: 10.3200/JOER.99.6.323-338

Shah, R., and Goldstein, S. M. (2006). Use of structural equation modeling in operations management research: looking back and forward. *J. Operat. Manag.* 24, 148–169. doi: 10.1016/j.jom.2005.05.001

Tabachnick, B. G. (2013). *Using Multivariate Statistics*. Boston: Pearson Education.

Tan, S.-H., and Guo, Y.-Y. (2008). Revision of self-control scale for Chinese college students. *Chin. J. Clin. Psychol.* 16, 468–470.

Tangney, J. P. (1991). Moral affect: the good, the bad, and the ugly. *J. Pers. Soc. Psychol.* 61, 598–607. doi: 10.1037/0022-3514.61.4.598

Tangney, J. P. (1995). Recent advances in the empirical study of shame and guilt. *Am. Behav. Sci.* 38, 1132–1145. doi: 10.1177/0002764295038008008

Tangney, J. P., Baumeister, R. F., and Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *J. Pers.* 72, 271–324. doi: 10.1111/j.0022-3506.2004.00263.x

Tangney, J. P., Hill-Barlow, D., Wagner, P. E., Marschall, D. E., Borenstein, J. K., Sanftner, J., et al. (1996). Assessing individual differences in constructive versus destructive responses to anger across the lifespan. *J. Pers. Soc. Psychol.* 70, 780–796. doi: 10.1037/0022-3514.70.4.780

Unger, A., Bi, C., Xiao, Y.-Y., and Ybarra, O. (2016). The revising of the Tangney Self-Control Scale for Chinese students. *PsyCh J.* 5, 101–116. doi: 10.1002/pchj.128

Wang, A. (2002). Validation of a Self-Control Rating Scale in a Chinese preschool. *J. Res. Childh. Educ.* 16, 189–201. doi: 10.1080/02568540209594984

Weng, X., and Chui, W. H. (2018). Assessing two measurements of self-control for Juvenile delinquency in China. *J. Contemp. Crim. Justice* 34, 148–167. doi: 10.1177/1043986218761932

Zinbarg, R. E., Revelle, W., Yovel, I., and Li, W. (2005). Cronbach's alpha, Revelle's beta, and McDonald's (omega H) : their relations with each other and two alternative conceptualizations of reliability. *Psychometrika* 70, 123–133. doi: 10.1007/s11336-003-0974-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | Conceptualization of the dimensionality of the Brief Self-Control Scale.

		13-item BSCS#					13-item BSCS		8-item BSCS		10-item BSCS		11-item BSCS^			
		T1	**T2**	**T3**	**T4**	**T5**	**General self-discipline**	**Impulse control**	**Impulsivity**	**Restrain**	**Inhibition**	**Initiation**	**F1**	**F2**	**F3**	**F4**
BSCS1	I am good at resisting temptation		*					*		*	*			*		
BSCS2	I have a hard time breaking bad habits (reverse scored)			*			*			*	*				*	
BSCS3	I am lazy (reverse scored)			*			*		–	–		*			*	
BSCS4	I say inappropriate things (reverse scored)	*					*		–	–	–	–	*			
BSCS6	I do certain things that are bad for me, if they are fun (reverse scored)	*					*		*		*		*			
BSCS13	I refuse things that are bad for me	*					*			*	–	–	–	–	–	–
BSCS17	I wish I had more self-discipline (reverse scored)					*	*		*		*		–	–	–	–
BSCS22	people would say that I have iron self-discipline		*				*		–	–		*		*		
BSCS28	pleasure and fun sometimes keep me from getting work done (reverse scored)				*			*	–	–		*				*
BSCS29	I have trouble concentrating				*		*		*		*					*
BSCS30	I am able to work effectively toward long-term goals		*				*		*			*		*		
BSCS31	sometimes I can't stop myself from doing something, even if I know it is wrong (reverse scored)	*						*	–	–	*		*			
BSCS32	I often act without thinking through all the alternatives (reverse scored)	*						*		*	–	–	*			

Source: Ferrari et al. (2009), de Ridder et al. (2011), Maloney et al. (2012), Tangney et al. (2004). #T1, general capacity for self-discipline; T2, inclination toward deliberate/non-impulsive action; T3, healthy habits; T4, self-regulation in service for a work ethic; T5, reliability. ^F1, self-discipline; F2, impulsivity; F3, healthy habits; F4, self-regulation.

ORIGINAL RESEARCH
published: 28 January 2020
doi: 10.3389/fpsyg.2020.00029

Measurement Invariance of the Prosocial Behavior Scale in Three Hispanic Countries (Argentina, Spain, and Peru)

Manuel Martí-Vilar[1], *César Merino-Soto*[2]* *and Lucas Marcelo Rodriguez*[3]

[1] Departament de Psicologia Bàsica, Facultat de Psicologia, Universitat de València, Valencia, Spain, [2] Instituto de Investigación de Psicología, Universidad de San Martín de Porres, Lima, Peru, [3] Centre for Interdisciplinary Research in Values, Integration and Social Development, Pontifical Catholic University of Argentina, Buenos Aires, Argentina

In a growing context of multiculturalism, prosocial behavior is important to build effective social exchange and service orientation among university students. The present study investigates prosocial behavior from a psychometric approach, to obtain evidence of the internal structure of the prosocial behavior scale (PS), in 737 young people enrolled at universities in Argentina (207), Spain (310), and Peru (220). First, the clarity of the items was explored in the three countries; second, possible irrelevant patterns of response, such as the careless and extreme responses, were evaluated; third, the non-parametric Mokken methodology was applied to identify the basic properties of the scale score; fourth, the structural equation modeling (SEM) methodology was used to identify the properties of the internal structure (dimensionality, tau-equivalence) of the latent construct; fifth, the measurement invariance according to sex (intra-equivalence) and country (inter-equivalence) was examined with the SEM methodology and other complementary strategies. Finally, reliability and internal consistency were evaluated both at score level and at item level. Implications for use of the PS instrument are discussed.

Keywords: prosocial, measurement invariance, social behavior, intercultural, university students, validation, assessment

Edited by:
Elisa Pedroli,
Italian Institute for Auxology (IRCCS),
Italy

Reviewed by:
Roger Watson,
University of Hull, United Kingdom
Cosimo Tuena,
Italian Institute for Auxology (IRCCS),
Italy

****Correspondence:***
César Merino-Soto
sikayax@yahoo.com.ar;
cmerinos@usmp.pe

Specialty section:
This article was submitted to
Quantitative Psychology
and Measurement,
a section of the journal
Frontiers in Psychology

Received: *16 May 2019*
Accepted: *07 January 2020*
Published: *28 January 2020*

Citation:
Martí-Vilar M, Merino-Soto C and Rodriguez LM (2020) Measurement Invariance of the Prosocial Behavior Scale in Three Hispanic Countries (Argentina, Spain, and Peru). Front. Psychol. 11:29. doi: 10.3389/fpsyg.2020.00029

INTRODUCTION

Prosocial behavior includes those actions tending to help or benefit other people, irrespective of the intention to be pursued with this help. Such behavior is the result of multiple individual and situational factors including parental variables and empathic traits (Eisenberg and Fabes, 1998). It is understood as a tendency to give rise to actions, belonging to the sphere of habits, practices and social interactions, that are characterized by the beneficent effects they produce on another person (Caprara, 2005). Moreover, Roche (2010) argued that truly prosocial behavior consists of help given to other people or groups in the absence of extrinsic or material reward. There are several different types of actions that make up prosocial behavior, such as physical and verbal help, material giving, verbal comfort, confirmation and positive appreciation of the other, deep listening, empathy, and solidarity, as well as the expression of unity with others (Roche, 1999).

Research on prosociality in diverse cultures has increased over the last few decades (Murakami et al., 2016; Luengo et al., 2017; Rodriguez et al., 2017; Gerbino et al., 2018). This has allowed researchers to carry out several meta-analysis studies on prosociality (Malti and Krettenauer, 2013; Shariff et al., 2016; Mesurado et al., 2019b), that show the value of clinical and educational interventions in encouraging prosocial behavior. For example, based on their own meta-analysis, Mesurado et al. (2019b) concluded that intervention programs aimed at promoting prosocial behaviors showed moderate effectiveness, while intervention programs focused on the prevention of aggressive were highly effective.

Since the construct of prosociality implicates a wide range of different behaviors, its measurement distinguishes between indicators of global prosocial behavior and prosocial behavior expressed in specific situations (Carlo and Randall, 2002). Measures of global prosocial behavior are defined as measures that evaluate personal tendencies to exhibit a series of prosocial behaviors across diverse social contexts and for different motives. An example of this type of global measure is the Prosociality Scale of Caprara et al. (2005). These global measures tend to characterize certain people as prosocial, distinguishing them from others who are not. However, global measures have limited application in research, since they do not investigate possible moderators such as in-group and out-group effects on tendencies to help, among other contextual factors. In contrast, measures of prosocial behavior in specific situations can provide information about more tightly delimited conceptualizations of prosociality, as well as supporting the elaboration and intercorrelation of different types of prosocial behavior. One example of this point is research that distinguishes between different recipients of aid, in terms of measuring the prosociality directed toward relatives, friends and strangers in adolescent populations (Padilla-Walker and Christensen, 2011; Padilla-Walker et al., 2015; Mesurado et al., 2019a). Such specific measures see prosociality as a multidimensional construct, which can be a very beneficial approach when studying interactions between prosociality and other variables (Carlo and Randall, 2002). However, the usefulness of a global or specific approach to measuring prosociality is not intrinsic to the measure itself, but is conditioned by the purpose of its use in basic or applied research, or in professional practice.

Another example of global prosociality measures is the Prosociality Scale (PS; Caprara et al., 2005), which describes the individual variability of prosocial behavior as a stable attribute, and is designed for young adults. It consists of 16 items to answer on an ordinal scale of 5 options ranging from "never/almost never" to "always/almost always." Based on the original study by the instrument's authors (Caprara et al., 2005), we can distinguish psychometrically the items that provide high information (items 3, 5, 7, 8, 10, 12, and 13), moderate information (items 4, 6, and 9) and low information (items 1, 2, 11, 14, 15, and 16). The PS has had some international diffusion, with studies in various countries. For example, investigations have been conducted with Colombian adolescents using the reduced version of the scale (Luengo et al., 2017). Studies have also been carried out in Japan (Murakami et al., 2016), and in a sample of Argentinian adolescents (Rodriguez et al., 2017). In the latter study a confirmatory factor analysis arrived at a scale of two dimensions (*prosocial behavior* and *empathy and emotional support*) while reducing the number of items to 10, and achieving an internal consistency of α = 0.78. Cross-cultural work has also been carried out on samples of children from Colombia, Italy, Jordan, Kenya, the Philippines, Sweden, Thailand, and the United States (Pastorelli et al., 2016), although data on the reliability and validity of the Prosociality Scale instrument were not presented in that study. It is worth mentioning that the aforementioned studies were carried out on children and adolescents, an age range for which the scale of Caprara et al. (2005) was not specifically designed. Their results should therefore be interpreted with caution, and should not automatically be generalized to adult populations.

Since the Prosociality Scale is recognized internationally, it is of great scientific and practical interest to evaluate its psychometric characteristics and variance across diverse populations. Additionally, studies that use a version of the scale in Spanish are particularly valuable since they are moderately scarce compared to studies that a use a version in English. Indeed, a recent systematic review of measures of prosocial behavior (Martí-Vilar et al., 2019) reported that PS is among the measures with few validation studies carried out adults, but with excellent internal consistency. The relationship between the importance of the construct and the its measurement in adults does not seem to be isomorphic, since there are few validation studies of internal structure and correlation studies with other relevant constructs: except for a study by Rodriguez et al. (2017), this information is practically absent in the Ibero-American population. These authors performed a confirmatory factor analysis on a population of Argentinian adolescents. In their study, a 10-item model with two dimensions was obtained, namely prosocial behavior on the one hand and empathy and emotional support on the other. In turn, they analyzed the convergent validity of the instrument, obtaining significant correlations with some dimensions of the scale of prosocial tendencies produced by Carlo and Randall (2002).

Investigations that have used the Prosociality Scale have rarely addressed certain aspects that could help to understand its psychometric functioning. For example, the functioning of the items within a tau-equivalent model has not been analyzed; this property is a condition for the use of the reliability coefficient type (Graham, 2006; Trizano-Hermosilla and Alvarado, 2016), as well as for identifying the homogeneity of the representation of the content and interpretation of the score. In this sense, because the factor loads signify the strength with which the items are connected to (represent) the latent construct (Trizano-Hermosilla and Alvarado, 2016), the similarity or dissimilarity of factor loads can influence interpretation of the score. Therefore, different factor load patterns (e.g., item 1: 0.80, item 2: 0.50, item 3: 30, item 4: 0.30; compared with item 1: 30, item 2: 30, item 3: 50, item 4:0.80), may not lead to the same interpretation of the construct.

On the other hand, all studies that have used the Prosociality Scale (except Caprara et al., 2005) have applied linear models that included latent variables (in other words, structural equation modeling, or SEM); however, a deeper analysis of the instrument requires considering that the interpretation rests on the score observed, and therefore a non-parametric methodology that uses the observed score as the main reference for the adjustment of the items may be necessary, and a prerequisite for the application of parametric models such as linear SEM modeling (Dima, 2018). The sequential or joint application of several procedures to identify the psychometric properties of a measure can be better understood within a framework of sensitivity analysis, in which the results of various methods or modifications of the data are contrasted, in order to evaluate the eventual convergence. This has been especially applied in the investigation of equivalence of measures (Hambleton, 2006; Teresi et al., 2009) and adaptation of evidence (Dima, 2018). Finally, due to the different informative strength of each PS item (as found by Caprara et al., 2005), it is plausible that each item is differently sensitive to factors such as sex; in this sense, the differences between groups in the means can mask fine differences at the item level. More precisely, descriptive analysis at the item level is relevant because each unit represents an elementary behavior of the intended construct, and its statistical behavior can help to better understand this, and precede the use of advanced analyses (Dima, 2018). Additionally, due to the apparent tendency to use single-item scales in self-report and epidemiological investigations, information at the item level can contribute to more informed choices in such uses.

The aim of the present study was to evaluate the psychometric functioning of the Prosocial Conduct Questionnaire in a context of intercultural use, focused on university participants from three Spanish-speaking countries: Argentina, Spain, and Peru. Specifically, the central objective was to obtain evidence of the validity of the internal structure of the Prosocial Behavior Questionnaire in three Hispanic countries, through the exploration of scalability, dimensionality, invariance of measurement and reliability of internal consistency. The aspects evaluated in this study may be specific to their use in these countries, and are linked to the evidence on the internal structure of the scale, which is a key component for other sources of evidence of validity (Lewis, 2017). Dimensionality, invariance and reliability can be considered fundamental contributors to the valid interpretation of a score, and together define an instrument's internal structure (Rios and Wells, 2014); that is, the theoretically coherent relationship between the components of a measure that serve as a basis for the interpretation of the score (American Educational Research Association [AERA] et al., 2014). Accordingly, evidence of validity based on the internal structure is critical in conditioning other evidence of validity (Ziegler and Hagemann, 2015). In the present study, scalability was also evaluated as a property of the score for establishing ordinal differences between subjects based on their observed scores (Mokken, 1971; van Schuur, 2003; Smits et al., 2012). This aspect is not necessarily equal to the dimensionality of an instrument, and therefore must be evaluated in a complementary way (Smits et al., 2012), usually with the non-parametric approach of Mokken (1971). The equivalence or invariance of measurement, as well as the similarity of internal consistency, and the sex differences in the level of total score and individual item, were also considered. Apparently, this is the first study that tests the dimensionality and invariance of the Prosociality Scale in several Ibero-American countries, and thus represents an advance toward the global use of the instrument.

MATERIALS AND METHODS

Participants

The study population were adult university students of Psychology, residing in Spanish-speaking metropolitan cities. The collected sample comprised 737 subjects, from Spain (n = 310), Peru (n = 220), and Argentina (n = 207), 568 being female (77.2%, the rest were all male). The distribution of sexes across the three countries (Argentina: 176 women, 85.0%, Peru: 143 women, 65.3%; Spain: 249 women, 80.3%) was moderately similar (Shanon index, H_{male} = 0.451, H_{female} = 0.465). Although there were statistically significant differences in the sex distributions (Marascuilo and McSweeney method, Marascuilo and McSweeney, 1967) between Peru and Argentina on the one hand, and Spain and Argentina on the other, these were moderate (d = 0.63) and small (d = 0.45), respectively; and overall they were small (Cohen-$w_{adjusted}$ = 0.273, Sheskin, 2007). The academic semesters sampled were the first (138, 18.8%), second (105, 14.3%), third (188, 25.5%), fourth (208, 28.3%), and fifth (97, 13.2%) semesters.

The total age in the sample was: M = 21.42, SD = 4.11, Min = 16, Max = 53); between the samples (Argentina: M = 20.67, SD = 2.88; Spain: M = 21.66, SD = 4.35; Peru: M = 21.79, SD = 4.66), the differences were statistically significant (F[2,733] = 4.926, $p < 0.01$) but the effect size ($\omega^2 = 0.01$) was very small (Field, 2013). The differences between distribution of semesters in Peru and Spain (Kolmogorov–Smirnov D = 0.386, $p < 0.01$), and Peru and Argentina (Kolmogorov–Smirnov D = 0.433, $p < 0.01$) were statistically significant, while those for Spain and Argentina were not (Kolmogorov–Smirnov: D = 0.084, $p > 0.10$). But the practical significance of these differences, in terms of similarity of frequencies (overlap, PSR, Rom and Hwang, 1996) tended to be high: PSR Peru–Spain = 80.7%; PSR Peru–Argentina = 78.4%; PSR Spain Argentina = 95.8%. According to previous studies of the validation and substantive use of the instrument in the adult population (Murakami et al., 2016; Pastorelli et al., 2016; Luengo et al., 2017; Rodriguez et al., 2017), the various sub-samples of our participants were not differentiated from one another in relation to sampling (non-probabilistic), coverage (young adults), or main activity (university studies), and therefore they can be thought of as generally aligned.

Instruments

Demographic Sheet

A questionnaire was compiled to gather sociodemographic information, namely country, city, age, sex, level of studies, and academic semester.

Prosociality Scale (Caprara et al., 2005)

This is a self-report measure that quantifies prosociality as a stable attribute in the adult population. It consists of 16 ordinally scaled items each with five response options. The response instructions posit a generic and timeless context of prosocial behaviors. In relation to the internal consistency of the instrument, the original authors reported unidimensionality, a wide range of psychometric precision, internal validity of the items, and internal consistency of α = 0.91 (Caprara et al., 2005). The Spanish version used here come from Rodriguez et al. (2017) for the Argentinian population.

Procedure

Data Collection

The study was authorized by the Ethics Committee of the Universitat de València. Participants were contacted at universities in Argentina, Peru, and Spain. If they wished to participate in the research, they were sent a link to an electronic form, where they had to complete a process of informed consent to answer the questionnaires. The entire sample was collected online.

Analysis

The analysis was divided into analysis of irrelevant answers, descriptive analysis of item responses, content validity testing on the clarity of the items, scalability of the score and the items, dimensionality of the score, internal consistency of the reliability estimates, and invariance and measurement equivalence.

Inattentive/irrelevant responses to content

For the present study, inattentive and irrelevant responses were explored, because answering questionnaires through a web platform has generally been associated with this type of irrelevant response pattern (Johnson, 2005). To identify this problem, the distance D^2 (Mahalanobis, 1936) was used to identify subjects who behaved as multivariate outliers; and to confirm this identification, the variability of intra-individual response was examined (IRV; Dunn et al., 2018). Both are effective techniques for this type of problem (Meade and Craig, 2012) and were implemented using the *careless* program (Yentes and Wilhelm, 2018).

Descriptive information

Tests of normality related to symmetry (D'Agostino, 1970) and kurtosis (Bonett and Seier, 2002) were used, as well as descriptive statistics to identify the floor and ceiling of each item.

Content validity

This part of the analysis highlighted the clarity of the content. The version of questionnaire used as a baseline of content was validated by Rodriguez et al. (2017). An independent evaluation of the content carried out by the authors indicated that it was phrased without apparent local expressions, and seemed generalizable across the participating groups. However, as (a) Spanish speech is generally characterized by local variations in the use of some words, and (b) there may be discrepancies in assessing clarity between expert judges and the participants themselves (Merino-Soto, 2016), we first corroborated whether the phrasing of the items was clear to the participants. For this purpose, they were given a score clarification form for the items. Each participant read the instructions first, and then scored each item using an ordinal scale of five points, from *Not clear* (1) to *Completely clear* (5). The ratings were analyzed using the V coefficient (Aiken, 1980), and their asymmetric confidence interval was computed using the *ICAiken* program (Merino-Soto and Livia, 2009). This coefficient is often used in content validity studies, to quantify the convergence of qualifying judges between values of 0 (absence of consensus) to 1 (complete consensus). To compare the perceived clarity between the three groups (Argentina, Spain, and Peru), a confidence interval of the difference between the V coefficients was applied (Merino-Soto, 2018). Acceptable clarity was established when the score estimates and the lower limit of the interval were above or equal to 0.60 (Merino-Soto and Livia, 2009).

Non-parametric analysis of scalability

To evaluate the fundamental properties of the instrument scores (Brodin, 2014), regardless of the strong presumptions of the latent variable models, a non-parametric approach (Mokken, 1971) was used to analyze the ordinal items of the Prosociality Scale (Molenaar and Sijtsma, 1988). This approach examines the ability of a score to differentiate the ordinal rank of the subjects or items of a measure. Its results are a prerequisite for more demanding parametric approaches (Brodin, 2014; Dima, 2018). There are several useful guides for conducting the analysis with the Mokken approach (e.g., Stochl et al., 2012; Watson et al., 2012; Sijtsma and van der Ark, 2017; Palmgren et al., 2018), but all converge on examining three basic properties for the completion of the *monotonic homogeneity model* (MHM; Sijtsma and van der Ark, 2017): (a) scalability of the items, using the H coefficient (Loevinger, 1948); (b) local independence, in which the responses to the items are not mutually influenced, examined by three conditional association indices, $W^{(1)}$, $W^{(2)}$ and $W^{(3)}$ (Straat et al., 2016); and (c) *monoticity*, that is, the function of incremental relation between the item and the latent attribute, evaluated by comparing the current and expected number of violations of the monotonic model (Mokken, 1971). The adjustment to this model generally uses the CRIT statistic, a diagnostic of the quality of the scale constructed using the weighted sum of several evaluative indicators. The result is a count of violations of the model, which through either a lax (CRIT > 80; van Schuur, 2003) or demanding criterion (CRIT > 40; Molenaar and Sijtsma, 2000), allows the identification of an excess of violations of the model, which would suggest removing the item.

For the selection of items, the following criteria were applied: (1) the point estimate of the coefficient *H* should be at least equal to or greater than 0.40 in the total sample; (2) the point coefficient H should be in at least two countries, equal to or greater than 0.40; (3) the lower limit of the IC in 90%, should be greater than 0.35; (4) no coefficient, in its point estimate or its lower limit, should be less than 0.30. This analytical procedure was performed using the *mokken* program (van der Ark, 2012; R Core Team, 2018).

Dimensionality and equivalence/invariance

To strengthen the assessment of dimensionality, the structured equation modeling (SEM) methodology was applied to identify the final characteristics of dimensionality and measurement invariance. To examine the dimensionality, we used a robust estimator for categorical variables (Muthén, 1984), which adjusts the first and second moments of the χ^2 statistic (mean-and-variance-adjusted unweighted least squares, or WLSMV; Muthén et al., 1997). This method uses a probit link to define the functional relationship between the items and the construct, as well as polychoric correlations between the items and the thresholds estimation to derive more precise parameters (e.g., factor loading) when the distributional asymmetry is strong (Sass et al., 2014; Li, 2016a,b). Potential changes in the re-specification of the measurement and invariance model were detected by (a) the modification index, at the nominal level 0.05 (WLSMV-$\chi^2 > 3.840$), and (b) in statistical power (Saris et al., 2009). IM is also a means of assessing local independence within SEM modeling (Douglas et al., 1998).

The sensitivity of each item with respect to its relation with the construct was estimated by means of a measure equivalent to the signal-to-noise ratio (SNR), which is generally an informative measure of the quality of the item, based on two information components: item discrimination and "noise" (residual variance not relevant to the construct; Ferrando, 2012a,b; Ferrando and Lorenzo-Seva, 2013). The SNR was obtained by squared factor loading (λ^2) on 1-λ^2. This relationship is usually binding with the IRT model (Cheng et al., 2012; Ferrando and Lorenzo-Seva, 2013), and is generally part of the reliability estimation for identifying the maximum variability linked to the construct (Bacon et al., 1995; Hancock and Mueller, 2001).

The heterogeneity of factor loads was tested by adjusting to the tau-equivalent model, implemented with a robust procedure (Yuan and Zhang, 2012) in the *coefficientalpha* program (Zhang and Yuan, 2015). The adjustment of the SEM model was evaluated with several practical indexes and conventional cut points: $\geq$0.95 for CFI and TLI; $\leq$0.08 for SRMR (Ullman, 2001). Although RMSEA can be recommended in modeling with categorical variables (Hutchinson and Olmos, 1998), it was not used to decide the adjustment due to its poor performance in models with small degrees of freedom (Kenny et al., 2015; Taasoobshirazi and Wang, 2016).

Invariance/measurement equivalence

This procedure was carried out in two phases, which looked at intra-country and inter-country equivalence. The intra-country equivalence was investigated in relation to participant sex, controlling the variability of the attribute effect (measured by the total score); to reduce the effect of cells with a small number of subjects (due to the distribution), the observed conditioning score (total score) was segmented into quintiles. The analysis used was the non-parametric differential item functioning (DIF), implemented with contingency tables for ordinal variables. The partial gamma coefficient was used (γ^p; Schnohr et al., 2008), with effect levels defined as weak (>0.15), moderate (0.16–0.30), and strong (>0.31). For the purposes of this study, general interpretation suggestions were used for γ^p (e.g., >0.60 = strong, >0.30 = moderate, and $\leq$0.30 = weak; Healey, 2012). This DIF procedure was required to address the small sample size of the compared groups (Lai et al., 2005; Güller and Penfield, 2009).

After verifying the intra-country equivalence, we continued by analyzing the equivalence between countries, through a sequence of steps appropriate for categorical variables (Wu and Estabrook, 2016), starting with a successive implementation of restrictions on the parameters of the items. The configurational invariance was analyzed first, followed by the cumulative restriction of equal thresholds, then the factorial loads, and finally the residuals. The SEM analyses were carried out with the *lavaan* (Rosseel, 2012) and *semtools* programs (Jorgensen et al., 2018). Since there are still no clear options of fit criteria for index of modification in the comparison of three groups, a liberal criterion was used to reduce the probability of Type I error. In this sense, Rutkowski and Svetina (2013) proposed less restrictive criteria in the comparison of more than two groups (but specifically, $\geq$ 10): Δ_{CFI}, Δ_{TLI} and Δ_{RMSEA}, changes less than 0.02; these criteria are similar to those conducted in large-scale studies and comparing more than two groups (OECD, 2014). For comparison purposes, criteria applied to IM were also used between two groups (Chen, 2007): $\Delta_{\text{CFI}} \leq 0.10$ and $\Delta_{\text{TLI}} \leq 0.10$. The convergence of the adjustment indices suggested the decision of indices of modification (IM), but since CFI is optimal in the comparison of nested models (Cheung and Rensvold, 2002) and reduces the Type I error (Elosua, 2011), some doubt can be resolved by the observation of CFI.

Reliability

Reliability was estimated at the item level and the score of each subscale. Regarding the items, the attenuated corrected coefficient (Wanous and Reichers, 1996) was used, given its lower bias and computational ease (Zijlmans et al., 2018); the minimum acceptable value is around 0.30 (Zijlmans et al., 2017). At the level of score, coefficients congruent with the non-parametric model were used (MS coefficient; Molenaar and Sijtsma, 1988), along with linear SEM modeling with the coefficient ω (Green and Yang, 2009) and bootstrap confidence intervals (500 replications) through the *coefficientalpha* program (Zhang and Yuan, 2015). For comparison purposes, the coefficient α was also calculated.

RESULTS

Inattentive/Irrelevant Responses to Content

Applying the Mahalanobis distance measure (D^2_{Median} = 13.914, min = 1.469, Q3 = 19.898), one participant (Peruvian) was detected with the maximum distance (D^2 = 138.72), and was 1.92 greater than the subject with the shortest distance (D^2 = 72.09). Although the χ^2 value was lower than the critical value (gl = 16, Bonferroni-α = 0.05, n = 46.03), the individual variability (IRV coefficient) for this participant corresponded with the maximum value of individual deviation (IRV = 1887), and it was also consistent in the identification of D^2. To reduce the probability that the identified participant was a "positive" or "negative" influential case in the adjustment due to its magnitude compared with the rest of the participants (Pek and MacCallum, 2011), this participant was removed, leading to a total sample of 736 for the following analyses.

Clarity of the Items

Table 1 shows the results of the evaluation of item clarity, as part of the content validity analysis. The point estimate of the coefficients was universally over 0.70, and their asymmetric confidence intervals were predominantly over 0.60; this is a minimally acceptable level (Merino-Soto and Livia, 2009). The average clarity in each group showed similarity between Argentinian and Spanish students (about 0.82), while it was comparatively low in Peruvian students (below 0.80), but nonetheless still at a satisfactory level of perceived clarity. For some items, the lower limit of the IC was below 0.60 (item 5 in Spain, item 11 in Peru, and item 8 in the three groups). These items were reviewed by the authors, especially item 8, where the psychometric behavior was observed in order to determine the effect of this relatively low perceived clarity. In the comparison between groups (through confidence intervals of the difference, in agreement with Merino-Soto, 2018), the most frequent discrepancies occurred among Peruvian students (perceived lower clarity) compared to Spanish and Argentinians, but the point estimates and their intervals in Peruvians tended to be acceptable. The lower limit of the interval for several items was around 0.05, indicating that in the population the difference detected might be small. At this stage, it was concluded that the clarity of the instrument was essentially satisfactory in the three groups.

Descriptive Statistics of the Items

Table 2 shows the items were distributed asymmetrically, with the highest density in the high response options; in the total sample, the asymmetry coefficients ($\sqrt{b_1}$) varied between −0.210 (item 11) and −1.065 (item 10). The kurtosis ($b_2 - 3$) showed more variability, with positive and negative values, and between −0.496 (item 11) and 1.041 (item 2). Overall, the items showed moderate or strong departures from normality (D'Agostino-Pearson K^2 between 15.3 and 112.5, $p < 0.01$).

In relation to some demographic variables (sex and age), in the total sample the Spearman correlation between the items and age was around zero (between −0.06 and 0.064), and predominantly without statistical significance. In each group, this trend was similar (Argentina: median = 0.042; Peru: median = −0.039; Spain: median = −0.024). Regarding sex, Spearman correlations varied between 0.030 (item 9) and 0.189 (item 4, female > male), and in each country it was also predominantly close to zero in Peru (median = 0.032), but around 0.10 in Argentina (median = 0.118) and Spain (median = 0.159). Finally, due to the tendency of responses toward high scores, several items in each country showed a ceiling effect, such that the minimum response was frequently option 2 or 3, especially in Spain and Peru. To align the analysis of latent variables with the methodology for categorical variables, options 1 and 2 were therefore integrated on these items, leaving the rest unmodified.

Non-parametric Analysis

Scalability

Regarding scalability (**Table 3**), in the first iteration of the analysis several items showed H scores below 0.40 in the three countries, as well as low levels of scalability in their confidence intervals (items 2, 9, 11, 12, and 16); other items showed comparatively weak H in at least two countries (items 1, 4, and 14). These items thematically corresponded to behaviors of

TABLE 1 | Coefficients V: clarity of content between participants (Argentina, Spain, and Peru).

	Coefficients V (IC 90%)									**Confidence interval for differences in V (90%)**					
	Argentina (*n* = 23)			**Spain (*n* = 24)**			**Peru (*n* = 23)**			**Arg. − Spa.**		**Arg. − Peru**		**Spa. − Peru**	
	V	**L**	**U**	**V**	**L**	**U**	**V**	**L**	**U**	**L**	**U**	**L**	**U**	**L**	**U**
Ps1	0.880	0.813	0.925	0.875	0.809	0.920	0.837	0.765	0.891	−0.076	0.085	−0.043	0.128	−0.047	0.123
Ps2	0.935	0.879	0.966	0.918	0.859	0.953	0.805	0.729	0.864	−0.049	0.084	**0.049**	**0.212**	**0.030**	**0.197**
Ps3	0.935	0.879	0.966	0.938	0.884	0.967	0.857	0.787	0.907	−0.066	0.059	**0.003**	**0.155**	**0.007**	**0.157**
Ps4	0.913	0.852	0.95	0.855	0.786	0.904	0.773	0.693	0.836	−0.020	0.136	**0.052**	**0.228**	−0.011	0.176
Ps5	0.837	0.765	0.891	**0.668**	**0.585**	**0.741**	0.740	0.659	0.808	**0.066**	**0.268**	−0.002	0.194	−0.179	0.037
Ps6	0.880	0.813	0.925	0.885	0.821	0.928	0.805	0.729	0.864	−0.085	0.073	−0.014	0.163	−0.007	0.167
Ps7	0.750	0.669	0.816	0.720	0.639	0.789	0.728	0.645	0.797	−0.076	0.134	−0.084	0.128	−0.114	0.100
Ps8	**0.620**	**0.534**	**0.699**	**0.520**	**0.437**	**0.602**	**0.663**	**0.578**	**0.738**	−0.019	0.215	−0.157	0.073	−0.255	−0.025
Ps9	0.958	0.908	0.981	0.845	0.775	0.896	0.837	0.765	0.891	**0.042**	**0.187**	**0.047**	**0.197**	−0.080	0.096
Ps10	0.945	0.892	0.973	0.970	0.926	0.988	0.815	0.740	0.872	−0.081	0.027	**0.052**	**0.210**	**0.083**	**0.232**
Ps11	0.880	0.813	0.925	0.813	0.739	0.869	**0.675**	**0.591**	**0.749**	−0.020	0.154	**0.105**	**0.30**	**0.033**	**0.239**
Ps12	0.825	0.751	0.881	0.875	0.809	0.920	0.695	0.611	0.767	−0.137	0.037	**0.027**	**0.231**	**0.082**	**0.275**
Ps13	0.945	0.892	0.973	0.938	0.884	0.967	0.783	0.704	0.845	−0.053	0.068	**0.08**	**0.246**	**0.073**	**0.239**
Ps14	0.837	0.765	0.891	0.698	0.616	0.768	0.750	0.669	0.816	**0.039**	**0.237**	−0.011	0.184	−0.157	0.055
Ps15	0.958	0.908	0.981	0.885	0.821	0.928	0.815	0.740	0.872	**0.007**	**0.141**	**0.067**	**0.221**	−0.016	0.156
Ps16	0.958	0.908	0.981	0.918	0.859	0.953	0.847	0.776	0.899	−0.021	0.103	**0.039**	**0.186**	−0.008	0.150
Media	0.879	–	–	0.833	–	–	0.777	–	–	–	–	–	–	–	–

Arg., Argentina; Spa., Spain; bold values, point coefficients below 0.70, lower interval below 0.60, or statistically significant difference; L, lower interval; U, upper interval.

TABLE 2 | Statistical descriptive information of items.

	Peru						Spain						Argentina					
	M	***SD***	**Min**	**Max**	**Floor**	**Ceiling**	***M***	***SD***	**Min**	**Max**	**Floor**	**Ceiling**	***M***	***SD***	**Min**	**Max**	**Floor**	**Ceiling**
Ps1	3.941	0.81	1	5	0.50	22.40	4.255	0.72	2	5	1.00	41.00	3.961	0.86	1	5	0.00	28.50
Ps2	4.114	0.84	1	5	0.50	35.60	4.445	0.65	1	5	0.30	51.90	4.184	0.86	2	5	0.00	42.00
Ps3	4.105	0.80	2	5	0.00	32.40	4.432	0.69	2	5	0.00	53.50	4.280	0.81	2	5	0.00	47.80
Ps4	3.836	1.00	1	5	3.20	27.40	3.735	1.02	1	5	3.20	24.80	3.652	1.17	1	5	4.30	30.00
Ps5	3.804	0.96	1	5	1.80	26.50	4.248	0.80	1	5	0.30	43.90	3.792	1.04	1	5	3.40	27.50
Ps6	3.890	0.83	2	5	0.00	25.10	4.016	0.78	2	5	0.00	28.10	3.792	0.91	1	5	1.00	23.20
Ps7	3.658	0.91	1	5	2.30	16.40	3.600	0.81	1	5	0.60	11.90	3.304	0.96	1	5	3.90	10.10
Ps8	3.845	0.86	1	5	0.90	23.70	4.435	0.77	1	5	0.30	57.40	4.039	0.91	1	5	1.00	36.20
Ps9	3.982	0.75	2	5	0.00	24.20	4.226	0.69	2	5	0.00	36.50	4.179	0.87	1	5	0.50	44.00
Ps10	3.945	0.94	1	5	0.50	30.10	4.455	0.65	2	5	0.00	53.20	4.256	0.85	1	5	1.00	46.90
Ps11	3.233	1.05	1	5	4.10	11.00	3.448	0.93	1	5	1.90	12.30	3.338	1.11	1	5	5.80	16.90
Ps12	3.594	0.96	1	5	2.30	16.00	4.077	1.01	1	5	4.50	37.40	3.705	1.06	1	5	3.40	27.10
Ps13	3.877	0.82	2	5	0.00	23.70	4.165	0.74	1	5	0.60	34.20	3.932	0.87	1	5	0.00	27.50
Ps14	4.046	0.77	2	5	0.00	27.40	4.335	0.65	3	5	0.00	43.20	4.164	0.89	1	5	0.50	42.50
Ps15	4.000	0.81	1	5	0.50	28.60	4.410	0.69	2	5	0.00	51.00	4.116	0.87	1	5	1.00	37.70
Ps16	4.009	0.91	1	5	1.40	32.90	4.342	0.69	2	5	0.00	45.50	4.203	0.87	1	5	0.50	44.40

M, mean; SD, standard deviation; Min, minimum score; Max, maximal score.

sharing personal resources (2, 9, 11, and 14), taking another's perspective in situations of discomfort (i.e., empathy; 12 and 16), and comfort and willingness to give help to others (1 and 4). The items that were satisfactorily maintained according to the initial criteria were items 3, 5, 6, 7, 8, 13, and 15, whose contents were distributed over helping behaviors (3, 6, and 7), empathy (5 and 8), and giving supportive company to others (13 and 15). Although item 10 (interpreted as providing help through emotional comfort) partially met the initial criteria, it was not included in the resulting version so as not to overemphasize the "helping" component in the instrument score. In the left section of **Table 3**, the results of the final iteration are shown. The scalability coefficient for the scale was 0.50 in the countries, and around 0.50 for each item (except item 15 that tended to be a little lower, though still close to 0.50). All were statistically significant with an alpha of 0.05 (for the items, z between 28.88 and 33.83; for the total score, $z = 59.83$).

Local Independence

In the analysis of conditional association (not shown in **Table 3**), the indices $W^{(2)}$ and $W^{(3)}$ did not detect any violation of local independence. Violations were found for $W^{(1)}$ between item 8 and items 5 ($W^{(1)} = 12.191$), 9 ($W^{(1)} = 10.227$), 12 ($W^{(1)} = 12.485$) and 16 ($W^{(1)} = 10.124$), and between item 13 and items 12 ($W^{(1)} = 13.096$) and 16 ($W^{(1)} = 13.349$). To corroborate this, within the next dimensionality analysis the indices of modification were evaluated.

Monotony

Finally, no violation of monotony was detected in the version obtained from seven items (see left side of **Table 3**). Based on the results of the non-parametric analysis as a whole, the obtained version had the following characteristics: the scalability of the score in the total sample and in each country was greater than 0.50, and its population variability was greater than 0.48, while each item showed a moderately similar magnitude of scalability, but generally greater than 0.50.

Dimensionality and Equivalence/Invariance

Analysis of Dimensionality (SEM)

Because the Prosociality Scale was apparently designed as a congeneric one-dimensional measure (without restriction of statistical equality between its items), the evaluation of the adjustment started with this model. The adjustment of the congeneric model with the 16 complete items was satisfactory according to the practical indices measure (see **Table 4**, results of the full version). The analysis of the modification indices indicated that potential mis-specifications were inconsistent according to the criteria of statistical power and practical significance (Saris et al., 2009). Given the strength of the adjustment and some trivial mis-specifications, this model was initially retained without add re-specifications. Although all the factorial loadings were statistically significant ($z > 10.0$), they varied from 0.500 to 0.811, which related to a large amount of variance in the construct (between 0.250 and 0.658, respectively). This suggested a wide range of variability (levels of 0.40, 0.50, 0.60, and 0.80; Beauducel and Wittmann, 2005). The SNR for each item emphasized the difference between the factorial loads, varying from 0.333 to 1.992, suggesting that the information relevant to the represented construct could range between very weak and very strong.

According to the results of the non-parametric analysis, a second iteration of the confirmatory factor analysis (CFA) was conducted, and the results of the model adjustment are shown in **Table 4** (results of the reduced version). These indicate a satisfactory adjustment, which was practically similar in the

TABLE 3 | Results of Mokken non-parametric analysis (scalability and monoticity).

	Scalability (*H* coefficient, first iteration)								**Scalability (*H* coefficient, second iteration)**								**Monoticity (*n* = 736)**		
	Total sample (*N* = 736)		**Argentina (*N* = 207)**		**Spain (*N* = 310)**		**Peru (*N* = 219)**		**Total sample (*N* = 736)**		**Argentina (*N* = 207)**		**Spain (*N* = 310)**		**Peru (*N* = 219)**		**#vi**	**$\#z_{sig}$**	**CRIT**
	H	***se***	***H***	***se***	***H***	***se***	***H***	***se***	***H***	***se***	***H***	***se***	***H***	***se***	***H***	***se***			
Ps1	0.402	0.025	0.374	0.049	0.345	0.032	0.432	0.046											
Ps2	0.379	0.026	0.333	0.050	0.275	0.032	0.465	0.041											
Ps3	0.492	0.022	0.503	0.041	0.408	0.030	0.526	0.038	0.571	0.025	0.590	0.045	0.493	0.039	0.592	0.041	0	0	0
Ps4	0.361	0.025	0.400	0.042	0.295	0.033	0.40	0.046											
Ps5	0.432	0.026	0.407	0.053	0.393	0.033	0.430	0.047	0.524	0.028	0.506	0.058	0.520	0.039	0.484	0.050	0	0	0
Ps6	0.455	0.024	0.467	0.044	0.410	0.031	0.460	0.044	0.560	0.025	0.563	0.046	0.545	0.039	0.542	0.043	0	0	0
Ps7	0.458	0.024	0.443	0.046	0.397	0.029	0.532	0.040	0.557	0.027	0.560	0.049	0.513	0.038	0.588	0.047	0	0	0
Ps8	0.467	0.023	0.454	0.045	0.386	0.033	0.487	0.042	0.555	0.025	0.547	0.051	0.523	0.037	0.527	0.047	0	0	0
Ps9	0.388	0.026	0.368	0.050	0.283	0.027	0.486	0.043											
Ps10	0.443	0.025	0.414	0.052	0.368	0.031	0.472	0.041											
Ps11	0.333	0.024	0.349	0.042	0.260	0.037	0.363	0.044											
Ps12	0.343	0.029	0.398	0.047	0.180	0.048	0.405	0.047											
Ps13	0.498	0.021	0.509	0.038	0.447	0.031	0.496	0.037	0.575	0.023	0.577	0.043	0.585	0.035	0.520	0.04	0	0	0
Ps14	0.399	0.025	0.371	0.045	0.306	0.031	0.475	0.043											
Ps15	0.45	0.023	0.458	0.038	0.342	0.032	0.487	0.044	0.487	0.028	0.480	0.052	0.397	0.045	0.520	0.04	0	0	0
Ps16	0.348	0.028	0.293	0.052	0.260	0.033	0.432	0.047											
H	0.413	0.019	0.407	0.036	0.33	0.023	0.456	0.035	0.546	0.022	0.545	0.043	0.512	0.031	0.537	0.040			

se, H standard error; #vi, number of violations to monoticity; $\#z_{sig}$, number of statistically significant violations; CRIT, combined count of #vi y $\#z_{sig}$.

specific indices compared with the full version (Δ_{CFI} = 0.005, Δ_{TLI} = 0.001, Δ_{SRMR} = 0.004). The adjustment without the recategorized items was also satisfactory, WLSMV-χ^2 = 155.3 (gl = 14, $p < 0.01$; CFI = 0.985, TLI = 0.978, SRMR = 0.061). These results were superior to the adjustment criteria chosen. All factorial loads were greater than 0.60, varying between 0.675 and 0.822; the change of the loads compared with the loads of the full version varied between | 0.1%| and | 7.9%|, while the factor loading of items 5, 6, 7, and 8 showed a small increase (between 1.4 and 5.7%). The adjustment with the reclassified items was indistinguishable from the results obtained before recategorization of the items (see **Table 4**).

After the congeneric modeling, in the adjustment of the tau-equivalent model, the common factor load were estimated as 0.764 (h^2 = 0.583). The adjustment was WLSMV-χ^2 = 207.8 (gl = 20, $p < 0.01$), CFI = 0.980, TLI = 0.979, SRMR = 0.069, RMSEA = 0.113 (IC 90% = 0.099, 0.127). Although the statistical test of tau-equivalence (Yuan and Zhang, 2012) rejected the null hypothesis of accepting this model, the differences of this model versus the congeneric model can be considered trivial: Δ_{CFI} = 0.005, Δ_{TLI} = 0.001, Δ_{SRMR} = 0.008.

Equivalence and Measurement Invariance

The intra-country analysis (see left part of **Table 4**) found that, once we controlled the performance on the observed score for the number of statistical tests (Bonferroni adjustment, p = 0.007), the tendency of the partial gamma coefficients (γ^p) was essentially concentrated on the weak level ($\leq$0.30). The items detected by possible uniform DIF (3 in Peru, and 5 in Spain) were examined in their content, and it was established that there was no reason to recognize any potential sources of DIF; therefore at this stage they were dismissed. On the other hand, although there were variations in the magnitude of the γ coefficient (not shown here) across quintiles, the homogeneity of the coefficients in the quintiles was confirmed (H-χ^2 < 15.0, Bonferroni adjusted p = 0.007), suggesting absence of non-uniform DIF.

Regarding the invariance/equivalence between countries, the baseline (configurational) model, along with the remaining models that included cumulative constraints, showed that the compared parameters (factorial loads, thresholds and residuals) changed only trivially (**Table 5**). Considering the chosen criteria (Chen, 2007; Rutkowski and Svetina, 2013; OECD, 2014), the equality constraints for each level of invariance produced results that suggested no invariance, and therefore it was concluded that there was compliance with the invariance across the three levels evaluated.

Reliability

In the total sample, we obtained an ω of 0.865 (*SE* = 0.009; 95% CI = 0.844,0.880); while α was 0.864 (SE = 0.009; 95% CI = 0.847,0.880). For practical purposes the two were indistinguishable. Estimated for each country, in Argentina (ω = 0.870, *SE* = 0.018, 95% CI = 0.830,0.899), Peru (ω = 0.890, *SE* = 0.016, 95% CI = 0.831,0.894), and Spain (ω = 0.845, *SE* = 0.015, 95% CI = 0.811,0.869), the coefficients were very similar and the variation could be due to sampling error. The α coefficients for each country (respectively 0.869, 0.842, and 0.869) showed insubstantial differences with the estimates of

TABLE 4 | Dimensionality (CFA-SEM) and differential item functioning (DIF).

	Dimensionality (CFA – SEM)						Differential item functioning (DIF)						Item-score reliability			
	Full version (n = 736)			Short version (n = 736)			Peru (n = 219)		Spain (n = 310)		Argentina (n = 207)		Total	Peru	Spain	Argentina
	λ	h^2	SNR	λ	h^2	SNR	γ^p	H-χ^2	γ^p	H-χ^2	γ^p	H-χ^2				
Ps1	0.653	0.426	0.743													
Ps2	0.627	0.393	0.648													
Ps3	0.800	0.640	1.778	0.776	0.601	1.514	0.374**	4.47	0.231	1.87	0.241	1.63	0.527	0.548	0.437	0.579
Ps4	0.567	0.321	0.474													
Ps5	0.734	0.538	1.168	0.777	0.603	1.524	0.175	5.04	0.441**	10.66	−0.184	14.93	0.458	0.404	0.475	0.418
Ps6	0.715	0.511	1.046	0.757	0.573	1.342	0.051	13.26	−0.166	5.81	−0.225	13.47	0.499	0.548	0.453	0.513
Ps7	0.701	0.492	0.966	0.722	0.522	1.089	−0.175	8.07	−0.098	1.62	−0.355	1.36	0.462	0.556	0.427	0.521
Ps8	0.811	0.658	1.922	0.824	0.678	2.115	−0.217	4.91	0.152	4.12	0.378	8.68	0.527	0.516	0.490	0.513
Ps9	0.615	0.378	0.608													
Ps10	0.718	0.515	1.064													
Ps11	0.500	0.250	0.333													
Ps12	0.591	0.349	0.537													
Ps13	0.797	0.635	1.741	0.799	0.639	1.765	0.341	10.15	−0.040	1.48	0.233	2.49	0.552	0.474	0.577	0.569
Ps14	0.666	0.444	0.797													
Ps15	0.733	0.537	1.161	0.671	0.450	0.819	−0.004	0.38	0.163	2.45	0.097	1.51	0.379	0.452	0.255	0.370
Ps16	0.566	0.321	0.471													
χ^2 (gl)	646.382 (104)			150.672 (14)												
CFI	0.980			0.985												
TLI	0.977			0.977												
RMSEA	0.084			0.115												
SRMR	0.065			0.063												

*λ, factor loading; h^2, total variance; SNR, signal-to-noise ratio; χ^2, WLSMV stimator; H-χ^2, strata homogeneity test of quintile score; γ^p, gamma partial coefficient. **$p < 0.007$.*

ω. The item-level reliability showed consistently high results in Argentina (median = 0.513, min. = 0.370, max. = 0.578), Peru (median = 0.516, min. = 0.403, max. = 0.556) and Spain (median = 0.452, min. = 0.255, max. = 0.577), and was similar between all three countries. Across the sample as a whole, the results were acceptable (see lower left side of **Table 4**).

TABLE 5 | Results of between invariance/equivalence (countries).

Invariance steps	WLSMV-χ^2 (gl)	CFI	TLI	SRMR	Δ_{CFI}	Δ_{TLI}	Δ_{SRMR}
Configurational	186.421 (42)	0.985	0.977	0.075	−0.009	−0.005	0.012
Weak (Metric)	284.553 (54)	0.976	0.972	0.087	0.000	0.009	−0.01
Strong (Scalar)	311.929 (80)	0.976	0.981	0.077	−0.008	−0.003	0.010
Strict	404.714 (94)	0.968	0.978	0.087	−0.009	−0.005	0.012

Δ, differences between fit indices CFI, TLI, and SRMR.

DISCUSSION

The present study applied psychometric methodology and rational-theoretical evaluations to refine the Prosociality Scale constructed by Caprara et al. (2005) for adult populations. Given the cross-cultural context of this study, it was particularly challenging to show the invariance of the scale's psychometric properties, and to date this is the only attempt at a cross-cultural psychometric exploration of the scale across several Spanish-speaking countries.

When the items were examined, they were characterized as not being distributed normally, characteristically with negative asymmetry. Also, the answers were oriented toward high response options. This trend was similar among the three countries examined. Associations with age were predominantly distributed around zero, both in the total sample and within individual countries. In contrast, relationships with sex were predominantly small in Spain (women > men), between trivial and small in Argentina (women > men), and completely trivial (around zero) in Peru. Considering that the differences in functioning of the items were trivial with respect to the sex of the participants, this finding for some individual items could lead to future explorations of differences at the level of the total score, but due to the strong asymmetry in the sex distribution in our samples, it would be best to avoid overinterpreting these results.

The fundamental psychometric criteria of our study were first based on a non-parametric method, created to evaluate the properties of measures that serve for ordering people based on their observed scores. Interestingly, the results of the application of the SEM and Mokken methodologies showed two things:

first, they tended to show convergence in the items with lower scalability and covariation with the construct, as identified in the study by Caprara et al. (2005); and second, items with comparatively poorer properties were more clearly identified by the non-parametric method (Mokken). Specifically, with the SEM method the items in general showed factor loads that are usually acceptable in the literature (>0.30 or >0.40), while these same levels applied to the *H* coefficient suggested a low scalability, and therefore lessened the discriminative ability of the observed score.

The content of the resulting scale was distributed over behaviors subsumed along one dimension, partially converging with the logic of another prosociality instrument created in one of the participating countries (Argentina), which is also applicable to university students (Auné et al., 2014). In that study, the instrument was multidimensional, with correlations between weak and moderate in the heterogeneous item-construct relationship (factorial loads). The two dimensions identified were interpreted as representing empathic behavior on the one hand, and initiative to help people on the other. In its analytical exploration, the former eigenvalue was very large in relation to the remaining values, and could suggest the exploration of a general latent variable, or that items with common variance load strongly toward a latent general factor. However, the difference between the one-dimensional model proposed here, and the multidimensional model proposed in the study of Auné et al. (2014) is influenced by the design of the theoretical constructions, and a combination of *post hoc* conceptual and empirical criteria to refine each instrument. Nevertheless, in our opinion the higher-order construct is prosocial behavior, supported by strongly intercorrelated specific content items. Thus, in the present study, conceptual decisions balanced purely empirical and mathematical decisions.

One of the evaluated characteristics was the adjustment to a tau-equivalent model (constraint of equality of factorial loads) compared with a congeneric model (in which factor loads were free to vary), which allowed us to identify the similarity in the construct representation of items and the appropriate reliability models. As in other Latin American studies (e.g., Auné et al., 2014, 2016), the heterogeneity of factor loads led to doubt about the appropriateness of internal consistency estimates such as the alpha coefficient, which assume the tau-equivalent model among the items. In the present study, although the statistical test of the difference between the congeneric and tau-equivalent models was statistically significant, the practical discrepancies between the two did not seem to be moderate or strong, but rather trivial. This leads to the conclusion that the items essentially showed similarity in their representativeness of the construct, and similar sensitivity to differentiate individual variability in the measured attributes. An additional advantage of adjusting the scale to a tau-equivalent model is that it helped to recover weak factorial models (Ximénez, 2006, 2016), and to avoid the rejection of models with salient factorial loading of 0.50 or less (Beauducel and Wittmann, 2005). Therefore, it is possible that the structure of the present version of the instrument can be replicated in future studies.

There are discrepancies in results regarding differences in prosociality according to participant sex (Martí-Vilar and Lorente, 2010). Some authors have argued that women show higher levels of prosociality, differences that are more marked in adult life (e.g., Eisenberg and Fabes, 1998). Other authors have noted that these sex differences depend on the motivation or type of prosocial behavior (Carlo et al., 2003; Auné et al., 2017). A plausible hypothesis that could explain this inconsistency is that certain instrument items but not others are psychometrically invariant. However, this was not verified in previous studies.

Although this study was carried out on a Spanish-speaking population, there are many differences between the societies of Spain, Argentina and Peru. Carballeira et al. (2014) showed that Latin American societies are more influenced by a collectivist culture, while Spanish society is more influenced by individualism. Such differences allow us to see the importance of this study since it involved testing the Prosociality Scale in countries with diverse cultural characteristics.

Due to the inconsistency of findings on the effect of sex on the variability of self-reported prosocial behavior, the investigation of equivalence was a preliminary, *sine qua non*, stage for the new version of the instrument. We found that, once the effect of the total score (measured as such) was controlled (using the DIF analysis approach), the differences in the answers were not outside the level of sampling error, and were generally trivial in magnitude. In the Peruvian and Spanish participants, two items worked differentially when the effect of the total score was controlled. Although the statistical detection of DIF does not directly indicate the absence of real bias (Lai et al., 2005), this is an avenue for further investigation. A qualitative analysis was beyond the objectives of this study, and thus the sources of this differential functioning were not qualitatively explored, so the conclusion of equivalence between men and women within each country is something to be tested by subsequent studies. Although this conclusion should be interpreted in the context of the limitations of the study (sample size and asymmetric proportion of men and women in each country), our results with the new reduced version can also be considered internally valid due to the strength of the unidimensional measurement model. As previously found, the unidimensionality of the new version is characterized by items with strong factorial loads, high signal-to-noise ratio, and an interdependent content relating to different observed behaviors.

Regarding the limitations of the study, one of these is the sample size. This can be considered large (>500) in terms of the total group size (Finch and French, 2008; Ximénez, 2016; Finch et al., 2018), but for the intra-country analysis it can be considered small (Ximénez, 2016). This could explain certain idiosyncratic variations between countries found in this sample. The intra-country sample sizes of our study, however, are typical of the common situation of small (or moderate) samples in social science research, and particularly in psychology (Beauducel and Wittmann, 2005). As more generally in psychology, the sample size of this study, in the total sample and in each subgroup, may generate suboptimal conditions for estimating psychometric parameters and their potential replicability. Although this problem is shared with many studies in the social sciences in general, and in psychology in particular (Beauducel and Wittmann, 2005), other aspects should also be considered to evaluate the potential replicability of our results: for example, the high magnitude of the factorial loading, as well as the convergence between the methodologies that were applied, and between the levels of statistical significance and practical significance that were found. Indeed, the application of several methods to identify dimensionality (within a framework of sensitivity analysis) can lead to more confidence in the results obtained, given the convergence observed.

A second limitation of the study is the asymmetric proportionality between men and women. However, the distribution of men and women in the sample may reflect current sex distributions among undergraduate students of psychology in Argentina, Spain, and Peru (and indeed other countries). Anecdotal evidence from the authors regarding said distribution supports this idea. A third limitation was the criterion used to decide on measurement invariance, since although the results of the adjustment met conventional criteria (≥0.90 or 0.95, Hu and Bentler, 1999) and other revised criteria (≥0.96; Yu, 2002), such criteria continue to be the subject of debate and further methodological research. This seems to be more prominent when comparing more than two groups (but less than ten), and in the context of asymmetric distribution of participants and moderately small sample size. The criteria applied in the present study (Chen, 2007; Rutkowski and Svetina, 2013; OECD, 2014) might produce Type I or II errors, and our criterion was essentially liberal. As the present study is one of the first of its kind, this decision should be re-evaluated for future studies. However, one aspect that balances this problem was that the approach of evaluating invariance/equivalence (applied to categorical variables) tends to yield robust and sensitive performance (Kim and Yoon, 2011; Sass et al., 2014). Another limitation is that the possible effect of the social desirability of the responses was not verified; this problem may have been reduced by the anonymity of data collection, or it may show correlations between moderate or weak (Rodrigues et al., 2017), and the reader is suggested to assess our results in the context of this limitation. Finally, other evidences of validity are required to corroborate the theoretical representation of this modified version of the instrument. Future studies should focus on the limitations of the study to advance the replicability of the results, as well as to obtain other evidence of validity required to open the way to substantive research with the instrument re-constructed here. This would contribute to our knowledge of prosociality measures, which are still an emerging area of investigation in measurement issues (Martí-Vilar et al., 2019).

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the United Nations Educational, Scientific and Cultural Organization (UNESCO), Declaration of Helsinki, and indicators of the Ethics Committee of the Universitat de València, No. H14820253925 (February 2, 2017). The studies involving human participants were reviewed and approved by the Ethic Committee of Universitat de Valéncia. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MM-V, CM-S, and LR designed the research and collected the data. CM-S analyzed the data. CM-S and LR interpreted the data. MM-V, CM-S, and LR drafted the manuscript. All authors critically revised the manuscript and gave their approval to the final version to be published.

ACKNOWLEDGMENTS

The authors thank the participating universities, as well as the participants, for the availability of facilities to develop and complete the study.

REFERENCES

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educ. Psychol. Measure.* 40, 955–959. doi: 10.1177/001316448004000419

American Educational Research Association [AERA], American Psychological Association [APA], and National Council for Measurement in Education [NCME] (2014). *The Standards for Educational and Psychological Testing.* Washington, D.C: AERA.

Auné, S., Abal, F., and Attorresi, H. (2014). Versión argentina de la escala de habilidades prosociales de Morales Rodríguez y Suárez Pérez (2011) [Argentine version of the Prosocial Skills Scale by Morales Rodríguez and Suárez Pérez (2011)]. *PRAXIS* 26, 31–48.

Auné, S., Abal, F., and Attorresi, H. (2017). Conducta prosocial y estereotipos de género [Prosocial behavior and gender stereotypes]. *PRAXIS* 27, 7–19.

Auné, S., Abal, J., and Attorresi, H. (2016). Diseño y construcción de una escala de conducta prosocial para adultos [Design and construction of a prosocial behavior scale for adults]. *Revista Iberoamericana de Diagnóstico y Evaluación y Evaluación Psicológica* 42, 15–25. doi: 10.21865/RIDEP42_15

Bacon, D. R., Sauer, P. L., and Young, M. (1995). Composite reliability in structural equation modeling. *Educ. Psychol. Measure.* 55, 394–406. doi: 10.1177/0013164495055003003

Beauducel, A., and Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Struc. Equa. Model.* 12, 41–75. doi: 10.1207/s15328007sem1201_3

Bonett, D. G., and Seier, E. (2002). A test of normality with high uniform power. *Comput. Stat. Data Anal.* 40, 435–445. doi: 10.1016/S0167-9473(02)00074-9

Brodin, U. A. (2014). *A '3 step' IRT Strategy for Evaluation of the Use of Sum Scores in Small Studies with Questionnaires Using Items with Ordered Response Levels.* Doctoral thesis, Karolinska Institutet, Stockholm.

Caprara, G. (2005). "Comportamento prosociale e prosocialità [Prosocial behavior and prosociality]," in *Il Comportamento Prosociale: Aspetti individuali, familiari e Sociali* [*Prosocial behavior: Individual, familiar and social aspects*>], eds G. V. Caprara, and S. Bonino, (Trento: Erikson), 7–22.

Caprara, G., Steca, P., Zelli, A., and Capanna, C. (2005). A new scale for measuring adults' prosocialness. *Eur. J. Psychol. Assess.* 21, 77–89. doi: 10.1027/1015-5759.21.2.77

Carballeira, M., González, J. A., and Marrero, R. J. (2014). Cross-cultural differences in subjective well-being: Mexico and Spain. *Ann. Psychol.* 31, 199–206. doi: 10.6018/analesps.31.1.166931

Carlo, G., Hausmann, A., Christiansen, S., and Randall, B. A. (2003). Sociocognitive and behavioral correlates of a measure of prosocial tendencies for adolescents. *J. Early Adolesc.* 23, 107–134. doi: 10.1177/0272431602239132

Carlo, G., and Randall, B. (2002). The development of a measure of prosocial behaviors for late adolescents. *J. Youth Adolesc.* 31, 31–44. doi: 10.1027/1015-5759.21.2.77

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struc. Equa. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Cheng, Y., Yuan, K.-H., and Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educ. Psychol. Measure.* 72, 52–67. doi: 10.1177/0013164411407315

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struc. Equa. Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

D'Agostino, R. B. (1970). Transformation to normality of the null distribution of G1. *Biometrika* 57, 679–681. doi: 10.2307/2334794

Dima, A. L. (2018). Scale validation in applied health research: tutorial for a 6-step R-based psychometrics protocol. *Health Psychol. Behav. Med.* 6, 136–161. doi: 10.1080/21642850.2018.1472602

Douglas, J., Kim, H., Habing, B., and Gao, F. (1998). Investigating local dependence with conditional covariance functions. *J. Educ. Behav. Stat.* 23, 129–151. doi: 10.2307/1165318

Dunn, A. M., Heggestad, E. D., Shanock, L. R., and Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: comparison to other indicators and relationships with individual differences. *J. Bus. Psychol.* 33, 105–121. doi: 10.1007/s10869-016-9479-0

Eisenberg, N., and Fabes, R. (1998). "Prosocial development," in *Handbook of Child Psychology, vol. 3: Social, Emotional, and Personality Development, Series*, eds W. Damon, and N. Eisenberg, (New York, NY: Wiley), 701–778.

Elosua, P. (2011). Assessing measurement equivalence in ordered-categorical data. *Psicológica* 32, 403–421.

Ferrando, P. J. (2012a). Assessing the discriminating power of item and test scores in the linear factor-analysis model. *Psicológica* 33, 111–134.

Ferrando, P. J. (2012b). Difficulty, discrimination, and information indices in the linear factor analysis model for continuous item responses. *Appl. Psychol. Measure.* 33, 9–24. doi: 10.1177/0146621608314608

Ferrando, P. J., and Lorenzo-Seva, U. (2013). *Unrestricted Item Factor Analysis and Some Relations With Item Response Theory*. Technical Report, Department of Psychology, Tarragona: Universitat Rovira i Virgili.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics*, 4th Edn. Los Angeles, CA: Sage.

Finch, H. W., French, B. F., and Hernández-Finch, M. E. (2018). Comparison of methods for factor invariance testing of a 1-factor model with small samples and skewed latent traits. *Front. Psychol.* 9:332. doi: 10.3389/fpsyg.2018.00332

Finch, W. H., and French, B. F. (2008). Comparing factor loadings in exploratory factor analysis: a new randomization test. *J. Modern Appl. Stat. Methods* 7:3. doi: 10.22237/jmasm/1225512120

Gerbino, M., Zuffianò, A., Eisenberg, N., Castellani, V., Luengo, B. P., Pastorelli, C., et al. (2018). Adolescents' prosocial behavior predicts good grades beyond intelligence and personality traits. *J. Pers.* 86, 247–260. doi: 10.1111/jopy.12309

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: what they are and how to use them. *Educ. Psychol. Measure.* 66, 930–944. doi: 10.1177/0013164406288165

Green, S. B., and Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: an alternative to coefficient alpha. *Psychometrika* 74, 155–167. doi: 10.1007/S11336-008-9099-3

Güller, N., and Penfield, R. D. (2009). A comparison of logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *J. Educ. Measure.* 46, 314–329. doi: 10.1111/j.1745-3984.2009.00083.x

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care* 11(Suppl. 3), S182–S188. doi: 10.1097/01.mlr.0000245443.86671.c4

Hancock, G. R., and Mueller, R. O. (2001). "Rethinking construct reliability within latent variable systems," in *Structural Equation Modeling: Present and Future—A Festschrift in Honor of Karl Jöreskog*, eds R. Cudeck, S. du Toit, and D. Soerbom, (Lincolnwood, IL: Scientific Software International), 195–216.

Healey, J. F. (2012). *The Essentials of Statistics: A Tool for Social Research*, 3rd Edn. Belmont, CA: Wadsworth.

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struc. Equa. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Hutchinson, S. R., and Olmos, A. (1998). Behavior of descriptive fit indexes in confirmatory factor analysis using ordered categorical data. *Struc. Equa. Model.* 5, 344–364. doi: 10.1080/10705519809540111

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *J. Res. Pers.* 39, 103–129. doi: 10.1016/j.jrp.2004.09.009

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2018). *semTools: Useful Tools for Structural Equation Modeling. R Package Version* 0.5-1. Available at: https://CRAN.R-project.org/package=semTools (accessed March 19, 2019).

Kenny, D. A., Kaniskan, B., and McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociol. Methods Res.* 44, 486–507. doi: 10.1177/0049124114543236

Kim, E. S., and Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Struc. Equa. Model.* 18, 212–228. doi: 10.1080/10705511.2011.557337

Lai, J. S., Teresi, J., and Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Eval. Health Profess.* 28, 283–294. doi: 10.1177/0163278705278276

Lewis, T. F. (2017). Evidence regarding the internal structure: confirmatory factor analysis. *Measure. Eval. Couns. Dev.* 50, 239–247. doi: 10.1080/07481756.2017.1336929

Li, C. H. (2016a). Confirmatory factor analysis with ordinal data: comparing robust maximum likelihood and diagonally weighted least squares. *Behav. Res. Methods* 48, 936–949. doi: 10.3758/s13428-015-0619-7

Li, C. H. (2016b). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychol. Methods* 21, 369–387. doi: 10.1037/met0000093

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychol. Bull.* 45, 507–530. doi: 10.1037/h0055827

Luengo, B. P., Eisenberg, N., Thartori, E., Pastorelli, C., Uribe, L. M., Gerbino, M., et al. (2017). Longitudinal relations among positivity, perceived positive school climate, and prosocial behavior in Colombian adolescents. *Child Dev.* 88, 1100–1114. doi: 10.1111/cdev.12863

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Natl. Instit. Sci.* 2, 49–55.

Malti, T., and Krettenauer, T. (2013). The relation of moral emotion attributions to prosocial and antisocial behavior: a meta-analysis. *Child Dev.* 84, 397–412. doi: 10.1111/j.1467-8624.2012.01851.x

Marascuilo, L. A., and McSweeney, M. (1967). Nonparametric post hoc comparisons for trend. *Psychol. Bull.* 67, 401–412. doi: 10.1037/h0020421

Martí-Vilar, M., Corell-García, L., and Merino-Soto, C. (2019). Systematic review of prosocial behavior measures. *Rev. Psicol.* 37, 349–377. doi: 10.18800/psico.201901.012

Martí-Vilar, M., and Lorente, S. (2010). "Determining factors of prosocial behavior," in *Moral Reasoning and Prosociality: Foundations*, ed. M. Martí-Vilar, (Madrid: Editorial CCS), 149–168.

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455. doi: 10.1037/a0028085

Merino-Soto, C. (2016). Percepción de la claridad de los ítems: comparación del juicio de estudiantes y jueces-expertos [Perception of item clarity: comparison of judgments between students and expertjudges]. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud* 14, 1469–1477. doi: 10.11600/1692715x.14239120615

Merino-Soto, C. (2018). Intervalos de confianza para la diferencia entre coeficientes de validez de contenido (V Aiken): sintaxis SPSS. [Confidence interval for difference between coefficients of content validity (Aiken's V): a SPSS syntax]. *Anal. Psicol.* 34, 587–590. doi: 10.6018/analesps.34.3.326801

Merino-Soto, C., and Livia, C. (2009). Intervalos de confianza asimétricos para el índice la validez de contenido: un programa Visual Basic para la V de Aiken. [Confidence interval for difference between coefficients of content validity (Aiken's V): a SPSS syntax]. *Anal. Psicol.* 25, 169–171. doi: 10.6018/analesps.34.3.283481

Mesurado, B., Guerra, P., De Sanctis, F., and Rodriguez, L. M. (2019a). Validation of the Spanish version of the prosocial behavior toward different targets scale. *Int. Soc. Work*. doi: 10.1177/0020872819858738

Mesurado, B., Guerra, P., Richaud, M. C., and Rodriguez, L. M. (2019b). "Effectiveness of prosocial behavior interventions: a meta-analysis," in *Psychiatry and Neuroscience Update*, eds P. Gargiulo, and H. Mesones Arroyo, (Cham: Springer), 259–271. doi: 10.1007/978-3-319-95360-1_21

Mokken, R. J. (1971). *A Theory and Procedure of Scale Analysis: With Applications in Political Research*. Berlin: De Gruyter Mouton.

Molenaar, I. W., and Sijtsma, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. *Kwantitatieve Methoden* 9, 115–126.

Molenaar, I. W., and Sijtsma, K. (2000). *MSP5 for Windows. A Program for Mokken Scale Analysis for Polytomous Items*. Groningen: ProGamma.

Murakami, T., Mishimura, T., and Sakurai, S. (2016). Prosocial behavior toward family, friends, and strangers: development of a prosocial behavior scale focused on the recipient of the behavior. *Jpn. J. Educ. Psychol.* 64, 156–169. doi: 10.5926/jjep.64.156

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika* 49, 115–132. doi: 10.1007/BF02294210

Muthén, B. O., du Toit, S. H. C., and Spisic, D. (1997). *Robust Inference Using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes*. Available at: https://www.statmodel.com/download/Article_075.pdf (accessed March 29, 2019).

OECD (2014). *TALIS 2013 Technical Report*. Paris: OECD Publishing.

Padilla-Walker, L. M., and Christensen, K. J. (2011). Empathy and self-regulation as mediators between parenting and adolescents' prosocial behavior toward strangers, friends, and family. *J. Res. Adolesc.* 21, 545–551. doi: 10.1111/j.1532-7795.2010.00695.x

Padilla-Walker, L. M., Dyer, W. J., Yorgason, J. B., Fraser, A. M., and Coyne, S. M. (2015). Adolescents' prosocial behavior toward family, friends, and strangers: a person-centered approach. *J. Res. Adolesc.* 25, 135–150. doi: 10.1111/jora.12102

Palmgren, P. J., Brodin, U., Nilsson, G. H. Watson, R. and Stenfers, T. (2018). Investigating psychometric properties and dimensional structure of an educational environment measure (DREEM) using Mokken scale analysis - a pragmatic approach. *BMC Med. Educ.* 18:235. doi: 10.1186/s12909-018-1334-8

Pastorelli, C., Lansford, J. E., Luengo, B., Malone, P. S., and Sorning, E. (2016). Positive parenting and children's prosocial behavior in eight countries. *J. Child Psychol. Psychiatry* 57, 824–834. doi: 10.1111/jcpp.12477

Pek, J., and MacCallum, R. C. (2011). Sensitivity analysis in structural equation models: cases and their influence. *Multivar. Behav. Res.* 46, 202–228. doi: 10.1080/00273171.2011.561068

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rios, J., and Wells, C. (2014). Validity evidence based on internal structure. *Psicothema* 26, 108–116. doi: 10.7334/psicothema2013.260

Roche, R. (1999). *Desarrollo de la inteligencia emocional y social desde los valores y Actitudes Prosociales en la escuela*>[*Development of social and emotional intelligence from prosocial values and attitudes in the school*>]. Buenos Aires: Ciudad Nueva.

Roche, R. (2010). *). Prosocialidad, nuevos desafíos*>[*Prosociality: New challenges*>]. Buenos Aires: Ciudad Nueva.

Rodrigues, J., Ulrich, N., Mussel, P., Carlo, G., and Hewig, J. (2017). Measuring prosocial tendencies in Germany: sources of validity and reliablity of the Revised Prosocial Tendency Measure. *Front. Psychol.* 8:2119. doi: 10.3389/fpsyg.2017.02119

Rodriguez, L. M., Mesurado, B., Oñate, M. E., Guerra, P., and Menghi, M. S. (2017). Adaptación de la Escala de prosocialidad de Caprara en adolescentes argentinos [adaptation of the prosociality scale of Caprara in argentinian adolescents]. *Rev. Eval.* 17, 177–187.

Rom, D. M., and Hwang, E. (1996). Testing for individual and population equivalence based on the proportion of similar responses. *Stat. Med.* 15, 1489–1505.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw*. 48, 1–36. doi: 10.3389/fpsyg.2014.01521

Rutkowski, L., and Svetina, D. (2013). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educ. Psychol. Measure.* 74, 31–57. doi: 10.1177/0013164413498257

Saris, W. E., Satorra, A., and van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Struc. Equa. Model.* 16, 561–582. doi: 10.1080/10705510903203433

Sass, D. A., Schmitt, T. A., and Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: a comparison of estimators. *Struc. Equa. Model.* 21, 167–180. doi: 10.1080/10705511.2014.882658

Schnohr, C. W., Kreiner, S., Due, E. P., Currie, C., Boyce, W., and Diderichsen, F. (2008). Differential item functioning of a family affluence scale: validation study on data from HBSC 2001/02. *Soc. Indic. Res.* 89, 79–95. doi: 10.1007/s11205-007-9221-4

Shariff, A. F., Willard, A. K., Andersen, T., and Norenzayan, A. (2016). Religious priming: a meta-analysis with a focus on prosociality. *Pers. Soc. Psychol. Rev.* 20, 27–48. doi: 10.1177/1088868314568811

Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*, 4th Edn. Boca Raton: Chapman and Hall.

Sijtsma, K., and van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *Br. J. Math. Stat. Psychol.* 70, 137–158. doi: 10.1111/bmsp.12078

Smits, I. A. M., Timmerman, M. E., and Meijer, R. R. (2012). Exploratory mokken scale analysis as a dimensionality assessment tool: why scalability does not imply unidimensionality. *Appl. Psychol. Measure.* 36, 516–539. doi: 10.1177/0146621612451050

Stochl, J., Jones, P. B., and Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Med. Res. Methodol.* 12:74. doi: 10.1186/1471-2288-12-74

Straat, J. H., van der Ark, L. A., and Sijtsma, K. (2016). Using conditional association to identify locally independent item sets. *Methodology* 12, 117–123. doi: 10.1027/1614-2241/a000115

Taasoobshirazi, G., and Wang, S. (2016). The performance of the SRMR, RMSEA, CFI, AND TLI: an examination of sample size, path size, and degrees of freedom. *J. Appl. Quant. Methods* 11, 31–39.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., et al. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): an item response theory approach. *Psychol. Sci. Q.* 51, 148–180.

Trizano-Hermosilla, I., and Alvarado, J. M. (2016). Best Alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Front. Psychol.* 7:769. doi: 10.3389/fpsyg.2016.00769

Ullman, J. B. (2001). "Structural equation modeling," in *Using Multivariate Statistics*, eds B. G. Tabachnick, and L. S. Fidell, (Boston: Allyn & Bacon), 653–771.

van der Ark, L. A. (2012). New developments in Mokken scale analysis in R. *J. Stat. Softw*. 48, 1–27. doi: 10.18637/jss.v048.i05

van Schuur, W. H. (2003). Mokken scale analysis: between the Guttman scale and parametric itemresponse theory. *Polit. Anal.* 11, 139–163. doi: 10.1093/pan/mpg002

Wanous, J. P., and Reichers, A. E. (1996). Estimating the reliability of a single-item measure. *Psychol. Rep.* 78, 631–634. doi: 10.2466/pr0.1996.78.2.631

Watson, R., van der Ark, L. A., Lin, L. C., Fieo, R., Deary, I. J., and Meijer, R. R. (2012). Item response theory: How Mokken scaling can be used in clinical practice. *J. Clin. Nurs.* 21, 2736–2746. doi: 10.1111/j.1365?2702.2011.03893.x

Wu, H., and Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika* 81, 1014–1045. doi: 10.1007/s11336-016-9506-0

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Struc. Equa. Model.* 13, 587–614. doi: 10.1207/s15328007sem1304_5

Ximénez, C. (2016). Recovery of weak factor loadings when adding the mean structure in confirmatory factor analysis: a simulation study. *Front. Psychol.* 6:1943. doi: 10.3389/fpsyg.2015.01943

Yentes, R. D., and Wilhelm, F. (2018). *Careless: Procedures for Computing Indices of Careless Responding. R packages version 1.1.0*. Available at: https://github.com/ryentes/careless (accessed April 16, 2019).

Yu, C. (2002). *Evaluating Cutoff Criteria of Model fit Indices for Latent Variable Models with Binary and Continuous Outcomes*. Unpublished dissertation,University of California, Los Angeles, CA.

Yuan, K.-H., and Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika* 77, 803–826. doi: 10.1007/s11336-012-9282-4

Zhang, Z., and Yuan, K. H. (2015). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: methods and software. *Educ. Psychol. Measure.* 76, 387–411. doi: 10.1177/0013164415594658

Ziegler, M., and Hagemann, D. (2015). Testing the unidimensionality of items: pitfalls and loopholes. *Eur. J. Psychol. Assess.* 31, 231–237. doi: 10.1027/1015-5759/a000309

Zijlmans, E. A. O., Tijmstra, J., van der Ark, L. A., and Sijtsma, K. (2017). Item-score reliability in empirical-data sets and its relationship with other item indices. *Educ. Psychol. Measure.* 78, 998–1020. doi: 10.1177/0013164417728358

Zijlmans, E. A. O., Van der Ark, L. A., Tijmstra, J., and Sijtsma, K. (2018). Methods for estimating item-score reliability. *Appl. Psychol. Measure.* 42, 553–570. doi: 10.1177/0146621618758290

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership