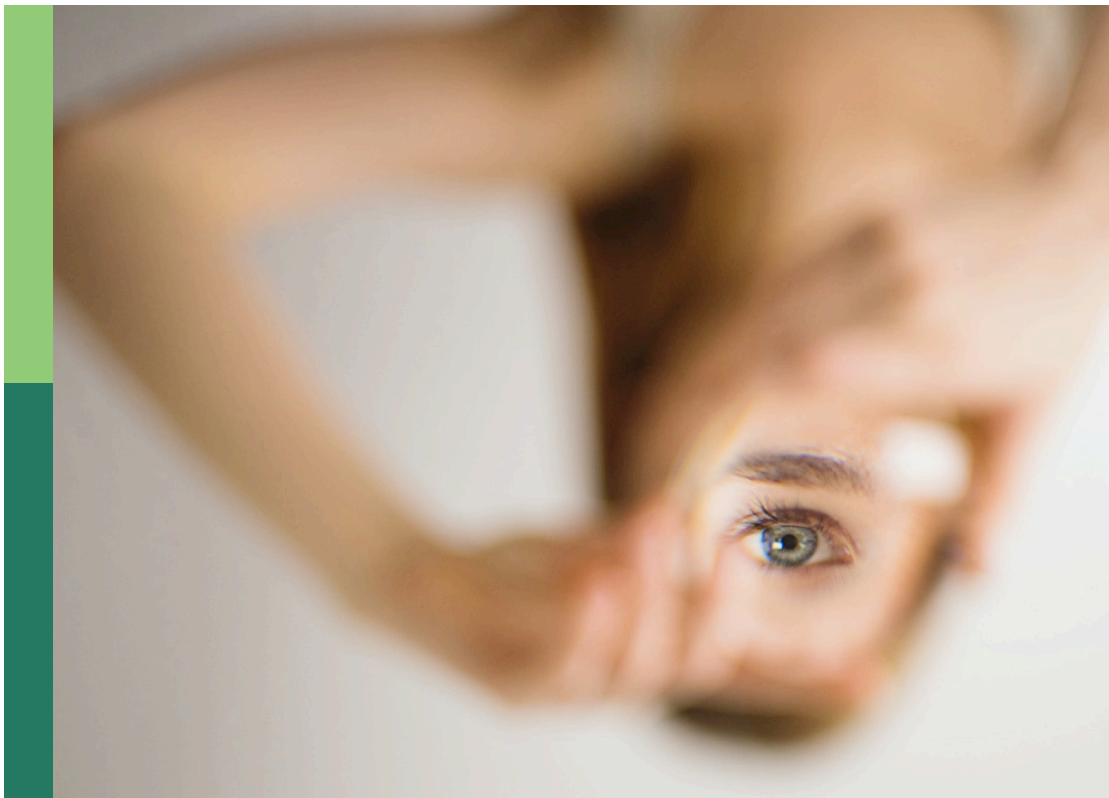
frontiers | Research Topics

New ideas in quantitative psychology and measurement

Edited by
Ioannis Tsaousis, Nikolaos Tsigilis, Georgios Sideridis and Iasonas Lamprianou

Published in
Frontiers in Psychology



ISSN 1664-8714
ISBN 978-2-8325-5421-0
DOI 10.3389/978-2-8325-5421-0

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: frontiersin.org/about/contact

New ideas in quantitative psychology and measurement

Topic editors

Ioannis Tsaousis — National and Kapodistrian University of Athens, Greece
Nikolaos Tsigilis — Aristotle University of Thessaloniki, Greece
Georgios Sideridis — Harvard Medical School, United States
Iasonas Lamprianou — University of Cyprus, Cyprus

Citation

Tsaousis, I., Tsigilis, N., Sideridis, G., Lamprianou, I., eds. (2024). *New ideas in quantitative psychology and measurement*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5421-0

Table of contents

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 30 August 2023
DOI 10.3389/fpsyg.2023.1248454

OPEN ACCESS

EDITED BY
Iasonas Lamprianou,
University of Cyprus, Cyprus

REVIEWED BY
Peida Zhan,
Zhejiang Normal University, China
George Karabatsos,
University of Illinois Chicago, United States

*CORRESPONDENCE
Jing Lu
✉ luj282@nenu.edu.cn
Guanghui Cheng
✉ chenggh845@nenu.edu.cn

RECEIVED 27 June 2023
ACCEPTED 10 August 2023
PUBLISHED 30 August 2023

CITATION
Wang X, Zhang J, Lu J, Cheng G and Shi N (2023) Exploration and analysis of a generalized one-parameter item response model with flexible link functions.
Front. Psychol. 14:1248454.
doi: 10.3389/fpsyg.2023.1248454

Exploration and analysis of a generalized one-parameter item response model with flexible link functions

Xue Wang[1], Jiwei Zhang[2], Jing Lu[1]*, Guanghui Cheng[3]* and Ningzhong Shi[1]

[1]Key Laboratory of Applied Statistics of Ministry of Education (MOE), School of Mathematics and Statistics, Northeast Normal University, Changchun, China, [2]Faculty of Education, Northeast Normal University, Changchun, China, [3]Guangzhou Institute of International Finance, Guangzhou University, Guangzhou, China

This paper primarily analyzes the one-parameter generalized logistic (1PGlogit) model, which is a generalized model containing other one-parameter item response theory (IRT) models. The essence of the 1PGlogit model is the introduction of a generalized link function that includes the probit, logit, and complementary log-log functions. By transforming different parameters, the 1PGlogit model can flexibly adjust the speed at which the item characteristic curve (ICC) approaches the upper and lower asymptote, breaking the previous constraints in one-parameter IRT models where the ICC curves were either all symmetric or all asymmetric. This allows for a more flexible way to fit data and achieve better fitting performance. We present three simulation studies, specifically designed to validate the accuracy of parameter estimation for a variety of one-parameter IRT models using the Stan program, illustrate the advantages of the 1PGlogit model over other one-parameter IRT models from a model fitting perspective, and demonstrate the effective fit of the 1PGlogit model with the three-parameter logistic (3PL) and four-parameter logistic (4PL) models. Finally, we demonstrate the good fitting performance of the 1PGlogit model through an analysis of real data.

KEYWORDS

Bayesian model evaluation criteria, item response theory, item characteristic curve, one-parameter generalized logistic models, STAN software

1. Introduction

Latent trait models, also known as item response theory (IRT) models, have gained widespread application in educational testing and psychological measurement (Lord and Novick, 1968; van der Linden and Hambleton, 1997; Embretson and Reise, 2000; Baker and Kim, 2004). These models utilize the probability of a response to establish the interaction between an examinee's "ability" and the characteristics of the test items, such as difficulty and guessing. The focus is on analyzing the pattern of responses rather than relying on composite or total score variables and linear regression theory. Specifically, IRT aims to model students' ability by examining their performance at the question level, providing a granular perspective on each student's ability based on the unique insights each question offers.

The Rasch model, also known as the one-parameter logistic IRT model, was innovated by Georg Rasch in 1960 and serves as a strategic tool in psychometrics for evaluating categorical data. This data includes responses to reading exams or survey questions and is analyzed in correlation with the trade-off between the respondent's ability, attitude, or personality trait and the item's difficulty (Rasch, 1960). For instance, this model could be used to determine a student's level of reading comprehension or gauge the intensity of

a person's stance on issues like capital punishment from their questionnaire responses. Beyond the realms of psychometrics and educational research, the Rasch model and its derivatives also find applications in diverse fields such as healthcare (Bezruczko, 2005), market research (Wright, 1977; Bechtel, 1985), and agriculture (Moral and Rebollo, 2017).

Within the framework of the Rasch model, the probability of a specific response–such as right or wrong–is modeled in relation to the examinee's ability and the item characteristic. Particularly, the classical Rasch model models the probability of a correct response as a logistic function of the discrepancy between the examinee's ability and the item difficulty. Typically, the model parameters depict the proficiency level of examinees and the complexity level of the items on a continuous latent scale. For instance, in educational assessments, the item parameter illustrates the difficulty level, whereas the person parameter represents the ability or attainment level of the examinee. The higher an individual's ability relative to the item difficulty, the higher the probability of a correct response. In cases where an individual's ability position equals the item difficulty level, the Rasch model inherently predicts a 50% chance of a correct response.

Parallel to the logistic IRT models, the normal ogive IRT models utilize the probit function to delineate the relationship between ability and item response, whereas the logistic IRT model employs the logit function to depict the same relationship. This constitutes a fundamental difference between the normal ogive IRT models and the more frequently utilized logistic IRT models. In fact, the use of the normal ogive model in the testing context has been further developed by a number of researchers. Lawley (1943, 1944) was the first to formally employ the normal ogive model to directly model binary item response data. Tucker (1946) used the term "item curve" to indicate the relationship between item response and ability. The early attempts at modeling binary response data culminated in the work of Lord (1952, 1953, 1980) who, unlike the early researchers, treated ability as a latent trait to be estimated and in doing so, laid the foundation for IRT.

The normal ogive IRT models (Lord, 1980; van der Linden and Hambleton, 1997; Embretson and Reise, 2000; Baker and Kim, 2004), also known as the one parameter normal ogive model, are a mathematical model used in the field of psychometrics to relate the latent ability of an examinee to the probability of a correct response on a test item. This model, as a component of IRT, facilitates the design, analysis, and scoring of tests, questionnaires, and comparable instruments intended for the measurement of abilities, attitudes, or other variables.

As previously noted, the Rasch model and the one-parameter normal ogive IRT model are premised upon symmetric functions to delineate the relationship between ability and item response, which result in a symmetric ICC. However, in certain contexts, these symmetric IRT models may not sufficiently capture the characteristics inherent in the data. These situations necessitate the utilization of asymmetric IRT models. Several asymmetric IRT models currently exist, such as the non-parametric Bayesian model, which constructs the ICC with a Dirichlet process prior (Qin, 1998; Duncan and MacEachern, 2008), and the Bayesian beta-mixture IRT model (BBM-IRT), which models the ICC with a flexible finite mixture of beta distribution (Arenson and Karabatsos, 2018). Karabatsos (2016) used the infinite mixture of normal c.d.f to model ICC, while Luzardo and Rodriguez (2015) constructed the ICC using the kernel regression method. There are also some skewed logistic IRT models, such as the logistic positive exponent (LPE) model and the reflection LPE (RLPE) model (Samejima, 1997, 1999, 2000; Bolfarine and Bazan, 2010; Zhang et al., 2022), which utilize skewed modifications of the logit links. Moreover, the positive trait item response model (PTIRM), which employs the log-logistic, lognormal, and Weibull as link functions to link the latent trait to the response, is used in some literature (Lucke, 2014; Magnus and Liu, 2018). In addition, the one-parameter complementary log-log IRT model also yields an asymmetric ICC (Goldstein, 1980; Shim et al., 2022). Compared to their symmetric counterparts, asymmetric IRT models can encapsulate a wider spectrum of data characteristics, particularly when the speed at which the probability of a positive response changes varies across different intervals of the latent trait. Furthermore, asymmetric IRT models are better suited to accommodate data where the probability of a positive response escalates more rapidly at higher trait levels and increases more sluggishly at lower trait levels. These asymmetric models, therefore, have a distinct advantage in capturing the nuanced dynamics of item responses that do not adhere strictly to symmetric patterns, thereby providing a more accurate representation of the interplay between individual ability and item response. As such, they represent a crucial development within the IRT field, broadening the applicability of these models in psychometric analyses and educational measurement.

This article discusses and analyzes the aforementioned one-parameter IRT models: the Rasch model, the one-parameter normal ogive IRT model, and the one-parameter complementary log-log IRT model. We propose a unified model representation that can encompass all three models through the manipulation of specific parameter values. In the present paper, our emphasis is placed on a class of generalized logistic models, introduced initially by Stukel (1988). This class of link functions is guided by a duo of parameters, precisely (η_1, η_2). By modulating the values of (η_1, η_2), this class is inclusive of logit, probit, complementary log-log link, along with an assortment of other symmetric and asymmetric links as particular instances. This class of models boasts sufficient versatility to accommodate the fitting of identical or diverse links to distinct items nested within the IRT model framework. An additional appealing characteristic of this class streamlines the execution of Markov chain Monte Carlo (MCMC) sampling from the posterior distribution via the recently formulated software, Stan. This research paper encompasses several key aspects. Firstly, we thoroughly discuss symmetric models such as the logit and probit models, as well as asymmetric models like the complementary log-log and generalized logit models, within the framework of a one-parameter IRT model. Secondly, we employ different links for different items in our analysis. Thirdly, we utilize the Stan platform to implement this flexible range of links for one parameter models and provide the corresponding Stan codes. By leveraging Stan, we are able to calculate deviance information criterion (DIC; Spiegelhalter et al., 2002) based on posterior distribution samples, which can naturally guide the selection of links and IRT model types. Lastly, through the 2015 computer-based PISA (Program for International Student Assessment)

sciences data, we empirically demonstrate that employing different generalized logit links for different items markedly improves data fit compared to traditional logistic, normal ogive and complementary log-log models, as determined by DIC criteria.

The remainder of this paper is organized as follows. In Section 2, we review the three one-parameter IRT models and the generalized logit link function, then introduce the main model of our study, namely the one-parameter generalized logistic (1PGlogit) model. In Section 3, we describe the Bayesian parameter estimation method that we use, discuss its software implementation, and elaborate on the Bayesian model assessment criteria we employ to evaluate the model fittings. Section 4 presents three simulation studies aimed at exploring the accuracy of model parameter estimation and assessing the fit of the 1PGlogit model in relation to various other symmetric or asymmetric models. In Section 5, we conduct an empirical study to validate the practical utility of the 1PGlogit model. Finally, in Section 6, we provide a summary of the paper.

2. Item response theory models with generalized logistic link functions

2.1. Overview of the one-parameter IRT models

The initial model in the field of IRT can be traced back to the 1930s, as proposed by Ferguson (1942), Lawley (1943), Mosier (1940, 1941), and Richardson (1936). It was later improved by Lord and Novick (1968) into what is now commonly referred to as the normal ogive model. Suppose we have N students each answering J items. Let X denote the response variable, and let x_{ij} be the response of the ith student ($i = 1, \cdots, N$) on the jth item ($j = 1, \cdots, J$). Here, $x_{ij} = 1$ indicates a correct answer, and $x_{ij} = 0$ indicates an incorrect one. Within the one-parameter normal ogive (1PNO) model, the probability of a correct response by the ith student on the jth item can be expressed as follows:

$$P(x_{ij} = 1|\theta_i, \beta_j) = \int_{-\infty}^{\theta_i - \beta_j} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz. \tag{1}$$

Here, β_j is the difficulty parameter of the jth item and θ_i is the latent trait of the ith student. A larger β_j implies a more difficult item, and the probability of a correct response increases with the increasing value of θ_i. As we can see, the 1PNO model is essentially a generalized linear model with a probit link.

Although the 1PNO model is quite interpretable and intuitive, its computation is complicated. In response to this, Rasch proposed the Rasch model in 1960, which was essentially a generalized linear model with a logit link. Specifically, the probability of a correct response in the model can be expressed in the following form:

$$P(x_{ij} = 1|\theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}, \tag{2}$$

where β_j and θ_i maintain the same interpretations as in the 1PNO model. In this form, to describe the probability of a student's response, it is no longer necessary to compute the cumbersome integrals, thereby simplifying the calculation.

Both of the models mentioned above possess a symmetrical item characteristic curve (ICC). However, Shim et al. (2022) proposed a one-parameter complementary log-log model (CLLM) which exhibits an asymmetric ICC. The probability of a correct response in the CLLM model can be expressed as follows:

$$P(x_{ij} = 1|\theta_i, \beta_j) = 1 - \exp\{-\exp(\theta_i - \beta_j)\}, \tag{3}$$

where β_j and θ_i retain the same interpretations as in the two models discussed earlier. As demonstrated by Shim et al. (2022), the CLLM possesses the capability to effectively address the guessing behavior exhibited by examinees in the three-parameter logistic (3PL) model and, in certain cases, can yield even better results. This implies that CLLM accounts for the effect of guessing. Essentially, the CLLM is a generalized linear model with a complementary log-log link.

2.2. Overview of the family of models based on generalized logit links

Let y be a dichotomous random variable. We assume that y equals 1 with probability $\mu(\eta)$ and 0 with probability $1 - \mu(\eta)$, where η is a linear predictor. Stukel (1988) introduced a class of generalized logistic models (Glogits), indexed by two shape parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$. Therefore, the Glogits model is controlled by a strictly increasing non-linear function $h_{\boldsymbol{\lambda}}(\eta)$. The specific expression is as follows:

$$\mu(\eta) = \frac{\exp\left\{h_{\boldsymbol{\lambda}}(\eta)\right\}}{1 + \exp\left\{h_{\boldsymbol{\lambda}}(\eta)\right\}}, \tag{4}$$

where the function $h_{\boldsymbol{\lambda}}(\eta)$ is defined as follows:
for $\eta > 0$ $\left(\mu(\eta) > \frac{1}{2}\right)$,

$$h_{\boldsymbol{\lambda}}(\eta) = \begin{cases} -\dfrac{\log(1 - \lambda_1 \eta)}{\lambda_1}, & \lambda_1 < 0, \\ \eta, & \lambda_1 = 0, \\ \dfrac{\exp(\lambda_1 \eta) - 1}{\lambda_1}, & \lambda_1 > 0. \end{cases} \tag{5}$$

for $\eta \leq 0$ $\left(\mu(\eta) \leq \frac{1}{2}\right)$,

$$h_{\boldsymbol{\lambda}}(\eta) = \begin{cases} \dfrac{\log(1 - \lambda_2 |\eta|)}{\lambda_2}, & \lambda_2 < 0, \\ \eta, & \lambda_2 = 0, \\ -\dfrac{\exp(\lambda_2 |\eta|) - 1}{\lambda_2}, & \lambda_2 > 0. \end{cases} \tag{6}$$

As evident from the above equations, the logit link serves as a special case of Glogits when $\lambda_1 = \lambda_2 = 0$. Furthermore, Stukel (1988) revealed that Glogits can be simplified to several other link functions under certain conditions. For instance, it reduces to a probit link when $\lambda_1 = \lambda_2 \approx 0.165$, a log-log link when $\lambda_1 \approx -0.037$ and $\lambda_2 \approx 0.62$, a complementary log-log link when $\lambda_1 \approx 0.62$ and $\lambda_2 \approx -0.037$, and a Laplace link when $\lambda_1 = \lambda_2 \approx -0.077$.

2.3. One-parameter generalized logistic IRT model

According to Glogit models, $\mu(\eta)$ forms a cumulative distribution function for η, which can be interpreted as the probability of a correct answer in IRT. Building on the traditional difficulty and ability parameters in a one-parameter IRT model, we reintroduce two shape parameters related to the item factors, denoted as $\boldsymbol{\lambda}_j = (\lambda_{1j}, \lambda_{2j})$. Consequently, we can deduce that the one-parameter generalized logistic model (1PGlogit) can be articulated as follows:

$$P(x_{ij} = 1|\theta_i, \beta_j, \boldsymbol{\lambda}_j) = \frac{\exp\left\{h_{\boldsymbol{\lambda}_j}\left(\theta_i - \beta_j\right)\right\}}{1 + \exp\left\{h_{\boldsymbol{\lambda}_j}\left(\theta_i - \beta_j\right)\right\}}. \tag{7}$$

Furthermore, when $\theta_i - \beta_j > 0$ (which implies that $P(x_{ij} = 1|\theta_i, \beta_j, \boldsymbol{\lambda}_j) > \frac{1}{2}$),

$$h_{\boldsymbol{\lambda}_j}\left(\theta_i - \beta_j\right) = \begin{cases} -\dfrac{\log(1 - \lambda_{1j}(\theta_i - \beta_j))}{\lambda_{1j}}, & \lambda_{1j} < 0, \\ \theta_i - \beta_j, & \lambda_{1j} = 0, \\ \dfrac{\exp\left(\lambda_{1j}(\theta_i - \beta_j)\right) - 1}{\lambda_{1j}}, & \lambda_{1j} > 0. \end{cases} \tag{8}$$

When $\theta_i - \beta_j \leq 0$, which implies that $(P(x_{ij} = 1|\theta_i, \beta_j, \boldsymbol{\lambda}_j) \leq \frac{1}{2})$,

$$h_{\boldsymbol{\lambda}_j}\left(\theta_i - \beta_j\right) = \begin{cases} \dfrac{\log(1 - \lambda_{2j}\left|\theta_i - \beta_j\right|)}{\lambda_{2j}}, & \lambda_{2j} < 0, \\ \theta_i - \beta_j, & \lambda_{2j} = 0, \\ -\dfrac{\exp\left(\lambda_{2j}\left|\theta_i - \beta_j\right|\right) - 1}{\lambda_{2j}}, & \lambda_{2j} > 0. \end{cases} \tag{9}$$

Specifically, when $\lambda_{1j} = \lambda_{2j} = 0$, the 1PGlogit model reduces to the Rasch model as shown in Equation (2); when $\lambda_{1j} = \lambda_{2j} \approx 0.165$, the 1PGlogit model becomes the traditional 1PNO model in Equation (1). This applies when $\theta_i - \beta_j \leq 0$, we have

$$P(x_{ij} = 1|\theta_i, \beta_j) = \frac{\exp\left\{-\dfrac{\exp\{0.165\left|\theta_i - \beta_j\right|\} - 1}{0.165}\right\}}{1 + \exp\left\{-\dfrac{\exp\{0.165\left|\theta_i - \beta_j\right|\} - 1}{0.165}\right\}}, \tag{10}$$

when $\theta_i - \beta_j > 0$, we have

$$P(x_{ij} = 1|\theta_i, \beta_j) = \frac{\exp\left\{\dfrac{\exp\{0.165(\theta_i - \beta_j)\} - 1}{0.165}\right\}}{1 + \exp\left\{\dfrac{\exp\{0.165(\theta_i - \beta_j)\} - 1}{0.165}\right\}}, \tag{11}$$

In fact, the CLLM model in Equation (3) is also a special case of the 1PGlogit model when the two shape parameters are restricted to $\lambda_{1j} \approx 0.62$ and $\lambda_{2j} \approx -0.037$. Specifically, when $\theta_i - \beta_j \leq 0$,

$$P(x_{ij} = 1|\theta_i, \beta_j) = \frac{\exp\left\{-\dfrac{\log\{1 + 0.037\left|\theta_i - \beta_j\right|\}}{0.037}\right\}}{1 + \exp\left\{-\dfrac{\log\{1 + 0.037\left|\theta_i - \beta_j\right|\}}{0.037}\right\}}, \tag{12}$$

when $\theta_i - \beta_j > 0$,

$$P(x_{ij} = 1|\theta_i, \beta_j) = \frac{\exp\left\{\dfrac{\exp\{0.62(\theta_i - \beta_j)\} - 1}{0.62}\right\}}{1 + \exp\left\{\dfrac{\exp\{0.62(\theta_i - \beta_j)\} - 1}{0.62}\right\}}, \tag{13}$$

To intuitively explore 1PGlogit IRT models, we visualize the ICCs of 1PGlogit IRT models with different λ_{1j} and λ_{2j} in Figure 1, where the difficulty parameter b is set as 0. It can be observed from Figure 1 that parameters λ_{1j} and λ_{2j} control the convergence speed of the tail of 1PGlogit. The speed at which the tail of the ICC approaches 0 can be referred to as the "rate of convergence to the lower limit". Similarly, the speed at which the ICC approaches 1 can be referred to as the "rate of convergence to the upper limit". Specifically, Figure 1A shows that the parameter λ_{1j} controls the convergence speed to the upper asymptote, while Figure 1B shows that the parameter λ_{2j} controls the convergence speed to the lower asymptote. Common to both parameters is that the larger the value of λ_{1j} (λ_{2j}), the faster the ICCs converge to the upper (lower) asymptote line. For instance, as shown in Figure 1A, when $\lambda_{1j} = 1$, the ICC of 1PGlogit(1, 0)has already converged to the upper asymptote $P(\theta) = 1$ before $\theta = 2$, while when $\lambda_{1j} = 0$, the ICC of 1PGlogit(0, 0) (i.e., Rasch model) just reaches the upper asymptote at $\theta = 4$. However, when $\lambda_{1j} = -1$, the ICC of 1PGlogit$(-1, 0)$ only converges to around $P(\theta) = 0.8$ at $\theta = 4$. The effect of the parameter λ_{2j} on the convergence of the ICC to the lower asymptote is similar to that of λ_{1j}, which can be seen in Figure 1B.

Based on the above analysis, it can be seen that the role of the parameter $\boldsymbol{\lambda}_j$ in 1PGlogit is somewhat analogous to the parameter c in the three-parameter logistic (3PL) model and the parameter d in the four-parameter logistic (4PL) model. As a result, we further compared the ICC of 1PGlogit with that of the 3PL model in Figure 2A and with the 4PL model in Figure 2B. Specifically, the expressions for the 3PL and 4PL models are as follows:

$$P(x_{ij} = 1|\theta_i, \alpha_j, \beta_j, c_j) = c_j + (1 - c_j)\frac{\exp\{\alpha_i(\theta_i - \beta_j)\}}{1 + \exp\{\alpha_i(\theta_i - \beta_j)\}}, \tag{14}$$

and

$$P(x_{ij} = 1|\theta_i, \alpha_j, \beta_j, c_j, d_j) = c_j + (d_j - c_j)\frac{\exp\{\alpha_i(\theta_i - \beta_j)\}}{1 + \exp\{\alpha_i(\theta_i - \beta_j)\}}. \tag{15}$$

In these models, α_j is the discrimination parameter, c_j is the lower asymptote parameter (which can be viewed as a guessing probability), and d_j is the upper asymptote parameter, where $1 - d_j$ can be considered as a slipping probability. For this analysis, we set $\alpha_j = 1$, $\beta_j = 0$, $c_j = 0.2$, and $d_j = 0.8$. As demonstrated in Figure 2A, the 3PL model has an upper asymptote at $P(\theta) = 1$ and a lower asymptote at $P(\theta) = 0.2$, while the 1PGlogit$(0, -1)$, with

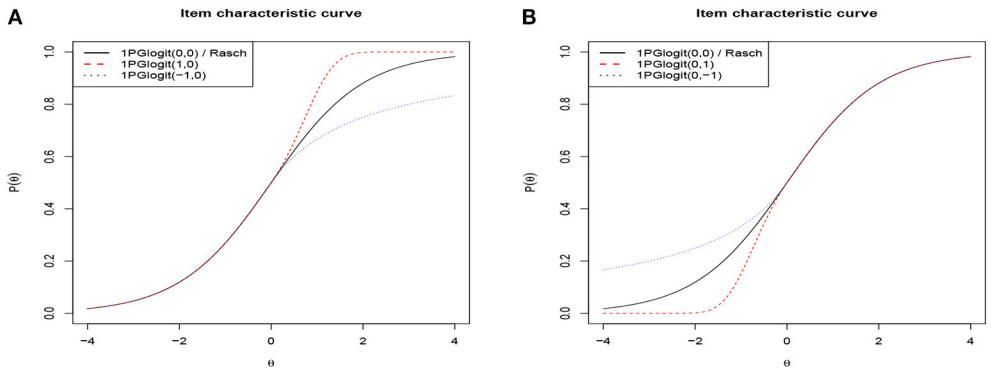


FIGURE 1
Item characteristic curves based on different the 1PGlogit models. **(A)** $\beta_j = 0$, $\lambda_{1j} = 0, 1, -1$ and $\lambda_{2j} = 0$. **(B)** $\beta_j = 0$, $\lambda_{1j} = 0$ and $\lambda_{2j} = 0, 1, -1$.

$\lambda_{1j} = 0$ and $\lambda_{2j} = -1$, displays an ICC similar to that of the 3PL model. In Figure 2B, the 4PL model exhibits an upper asymptote at $P(\theta) = 0.8$ and a lower asymptote at $P(\theta) = 0.2$. When $\lambda_{1j} = -1$ and $\lambda_{2j} = -1$, the 1PGlogit$(-1, -1)$ shows an ICC comparable to the 4PL model. Hence, the parameter $\boldsymbol{\lambda}_j$ in 1PGlogit can be adjusted to represent the assumed guessing and slipping behaviors in the 3PL and 4PL models.

3. Bayesian estimation and model evaluations

In this study, we adopt the Bayesian statistical inference method to estimate the parameters in 1PGlogit IRT models. Let $P_{ij} = p(x_{ij} = 1|\beta_j, \lambda_{1j}, \lambda_{2j}, \theta_i)$, which is defined as shown in Equations (7)–(9). Thus, the likelihood function for the response of the ith examinee to the jth item can be written as:

$$p(x_{ij}|\beta_j, \lambda_{1j}, \lambda_{2j}, \theta_i) = P_{ij}^{x_{ij}}(1 - P_{ij})^{1-x_{ij}}. \tag{16}$$

Let $\boldsymbol{x} = (x_i, \cdots, x_N)$, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_J)$, $\boldsymbol{\lambda_1} = (\lambda_{11}, \cdots, \lambda_{1J})$, $\boldsymbol{\lambda_2} = (\lambda_{21}, \cdots, \lambda_{2J})$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_N)$. Then the joint posterior distribution of parameters $\boldsymbol{\beta}, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}$, and $\boldsymbol{\theta}$ can be derived as:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\theta}|\boldsymbol{x}) &= p(\boldsymbol{x}|\boldsymbol{\beta}, \boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\theta}) p(\boldsymbol{\beta}) p(\boldsymbol{\lambda_1}) p(\boldsymbol{\lambda_2}) p(\boldsymbol{\theta}), \\ &= \underbrace{\left\{\prod_{i=1}^{N}\prod_{j=1}^{J} p(x_{ij}|\beta_j, \lambda_{1j}, \lambda_{2j}, \theta_i)\right\}}_{\text{Likelihood function}} \underbrace{\left\{\prod_{j=1}^{J} p(\beta_j) p(\lambda_{1j}) p(\lambda_{2j})\right\} \left\{\prod_{i=1}^{N} p(\theta_i)\right\}}_{\text{Prior distributions}}. \end{aligned} \tag{17}$$

3.1. Prior distributions

According to Chen et al. (2002) and Chen et al. (1999), it is necessary to constrain the parameters λ_{1j} and λ_{2j} to be greater than -1 to ensure a proper posterior distribution. Therefore, the priors for λ_{1j} and λ_{2j} should be truncated at -1. The parameters β_j and θ_i are assumed to follow different normal prior distributions, while λ_{1j} and λ_{2j} are assumed to follow a truncated normal prior distribution. Overall, the priors for the parameters are set as follows:

$$\begin{aligned} \beta_j &\sim \mathrm{N}(0, \sigma_\beta^2), \\ \lambda_{1j} &\sim \mathrm{N}(0, \sigma_\lambda^2)\mathcal{I}(-1, \infty), \\ \lambda_{2j} &\sim \mathrm{N}(0, \sigma_\lambda^2)\mathcal{I}(-1, \infty), \\ \theta_i &\sim \mathrm{N}(0, 1), \\ \sigma_\beta &\sim \mathrm{Cauchy}(0, 5)\mathcal{I}(0, \infty), \\ \sigma_\lambda &\sim \mathrm{Cauchy}(0, 5)\mathcal{I}(0, \infty), \end{aligned} \tag{18}$$

where $\mathcal{I}(a, b)$ implies that the parameter is constrained within the interval (a, b).

3.2. Stan software

In this paper, we employ the MCMC method for parameter estimation. Currently, there are various software options available for implementing the MCMC algorithm, such as WinBUGS (Lunn et al., 2000), OpenBUGS (Spiegelhalter et al., 2010), and JAGS (Plummer, 2003). However, In the subsequent research, we utilize the Stan software (Stan Development Team, 2019), which is based on the Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011) and the no-U-turn sampler (NUTS) (Hoffman and Gelman, 2014). HMC efficiently explores posteriors in models and is often faster than the Gibbs method (Geman and Geman, 1984) and the Metropolis algorithm (Metropolis et al., 1953), while NUTS further improves efficiency. Additionally, Stan provides interfaces with data analysis languages such as R, Python, Matlab, etc., making

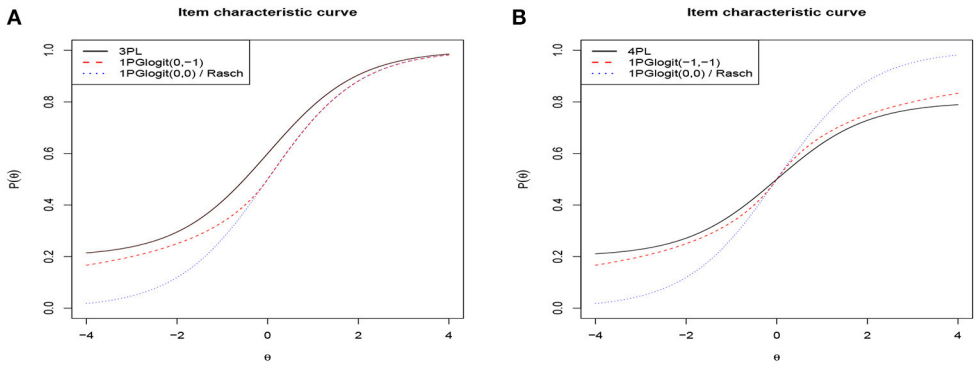


FIGURE 2
Item characteristic curves based on 3PL, 4PL, and 1PGlogit models. **(A)** 3PL model with $\alpha_j = 1$, $\beta_j = 0$, $c_j = 0.2$. **(B)** 4PL model with $\alpha_j = 1$, $\beta_j = 0$, $c_j = 0.2$, $d_j = 0.8$.

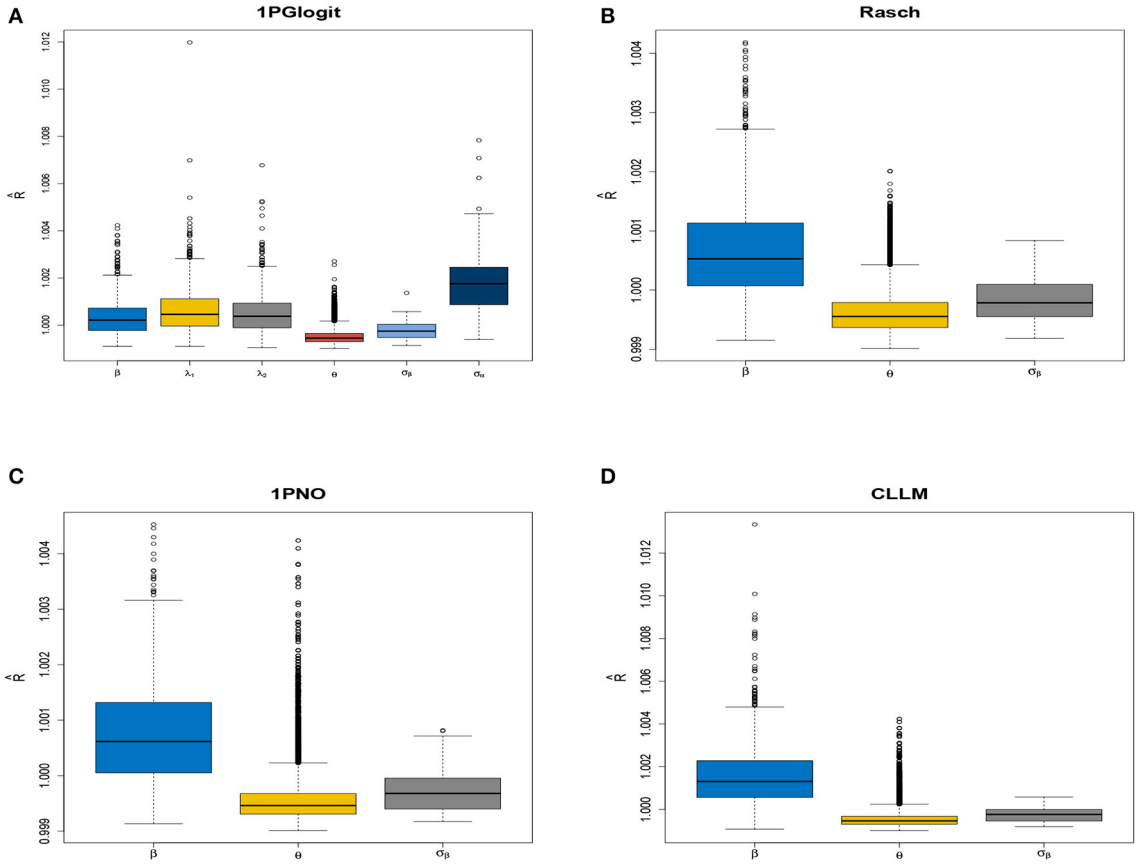


FIGURE 3
Boxplot of parameter $\hat{R}$ in four models under $N = 1,000$ and $J = 20$ conditions in simulation 1. **(A)** 1PGlogit model. **(B)** Rasch model. **(C)** 1PNO model. **(D)** CLLM.

it convenient for our use. To implement the Stan program, we specifically utilize the R package **rstan**, which interfaces with Stan in R (R Core Team, 2019). The Stan code employed for parameter estimation in this study, along with the actual data, can be found at the following URL: https://github.com/X-Wang777/-A-Generalized-One-Parameter-IRT. Furthermore, Luo and Jiao (2018) offer a detailed tutorial on utilizing Stan for estimating various IRT models.

10.3389/fpsyg.2023.1248454

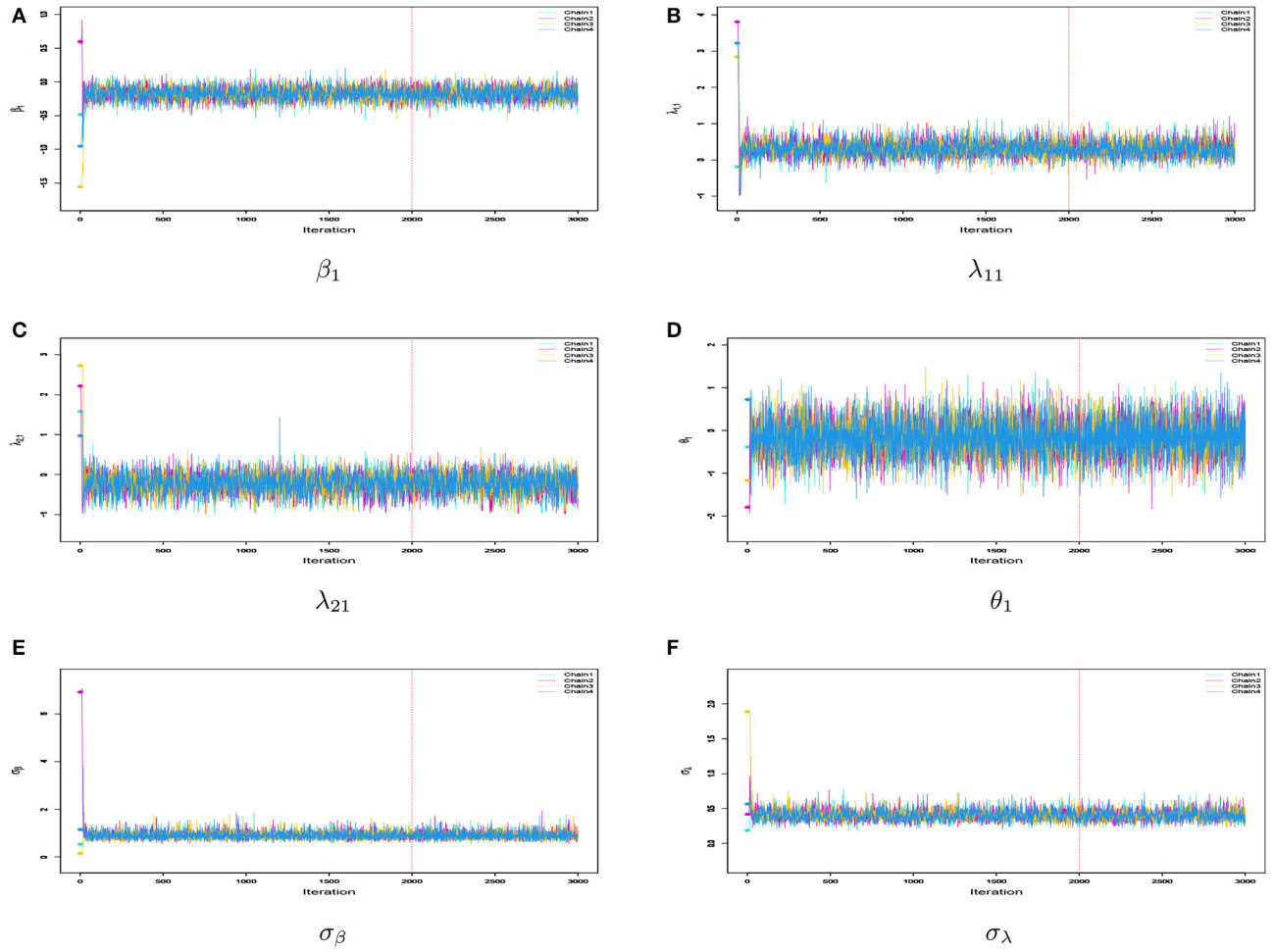


FIGURE 4
(A–F) Sampling trace plots of parameters β_1, λ_{11}, λ_{21}, θ_1, σ_β, and σ_λ in four Markov chains for 1PGlogit model under $N = 1,000$ and $J = 20$ conditions in simulation 1.

3.3. Criteria for assessing parameter estimation accuracy

In this research, we will use four criteria for assessing the accuracy of parameter estimation. They are Bias, RMSE (Root Mean Squared Error), SE (Standard Error), and SD (Standard Deviation). Assuming the parameter of interest is β_j, the evaluation criteria based on the β_j parameter are defined as follows:

$$\text{Bias}(\beta_j) = \frac{1}{R}\sum_{r=1}^{R}(\widehat{\beta}_j^{(r)} - \beta_j),$$

$$\text{RMSE}(\beta_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\beta}_j^{(r)} - \beta_j)^2},$$

$$\text{SE}(\beta_j) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\beta}_j^{(r)} - \frac{1}{R}\sum_{l=1}^{R}\widehat{\beta}_j^{(l)}\right)^2},$$

$$\text{SD}(\beta_j) = \frac{1}{R}\sum_{r=1}^{R} SD^{(r)}(\beta_j).$$

where R denotes the number of replications and $\widehat{\beta}_j^{(r)}$ is the estimate of β_j in the rth replication, and $\text{SD}^{(r)}(\beta_j)$ is the posterior standard deviation of β_j in the rth replication. Thus, we are able to calculate the average values for the four accuracy assessment indicators based on all items. That is,

$$\text{Average Bias}(\boldsymbol{\beta}) = \frac{1}{J\times R}\sum_{j=1}^{J}\sum_{r=1}^{R}(\widehat{\beta}_j^{(r)} - \beta_j),$$

$$\text{Average RMSE}(\boldsymbol{\beta}) = \frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\beta}_j^{(r)} - \beta_j)^2},$$

$$\text{Average SE}(\boldsymbol{\beta}) = \frac{1}{J}\sum_{j=1}^{J}\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\beta}_j^{(r)} - \frac{1}{R}\sum_{l=1}^{R}\widehat{\beta}_j^{(l)}\right)^2},$$

$$\text{Average SD}(\boldsymbol{\beta}) = \frac{1}{J\times R}\sum_{j=1}^{J}\sum_{r=1}^{R} SD^{(r)}(\beta_j).$$

TABLE 1 Evaluating the accuracy of parameter estimation for various models and simulation conditions in simulation study 1.

		$N = 1,000$				$N = 2,000$			
		Bias	*MSE*	*SE*	*SD*	*Bias*	*MSE*	*SE*	*SD*
$J = 20$									
1PGlogit	β	−0.0310	0.0102	0.0869	0.1009	−0.0361	0.0063	0.0642	0.0729
	λ_1	−0.0452	0.0972	0.1381	0.2246	0.0123	0.0566	0.1359	0.1945
	λ_2	−0.0406	0.0785	0.1513	0.2530	−0.0284	0.0508	0.1395	0.2141
	θ	−0.0281	0.1869	0.3825	0.4416	0.0389	0.1907	0.3803	0.4311
Rasch	β	0.0329	0.0067	0.0728	0.0792	−0.0033	0.0027	0.0498	0.0563
	θ	0.0337	0.2058	0.3988	0.4583	−0.0024	0.2168	0.3993	0.4597
1PNO	β	0.0287	0.0051	0.0642	0.0740	0.0166	0.0025	0.0463	0.0526
	θ	0.0344	0.1879	0.3807	0.4293	0.0148	0.1857	0.3809	0.4293
CLLM	β	−0.0146	0 .0046	0.0639	0.0710	-0.0095	0.0020	0.0428	0.0505
	θ	−0.0103	0.1701	0.3616	0.4073	−0.0058	0.1697	0.3608	0.4082
$J = 40$									
1PGlogit	β	−0.0434	0.0138	0.0936	0.1082	0.0064	0.0067	0.0728	0.0798
	λ_1	−0.0383	0.0912	0.1503	0.2449	−0.0201	0.0631	0.1344	0.2079
	λ_2	0.0318	0.0938	0.1755	0.2574	−0.0167	0.0684	0.1575	0.2111
	θ	−0.0280	0.1166	0.3102	0.3378	−0.0091	0.1135	0.3104	0.3373
Rasch	β	0.0296	0.0061	0.0713	0.0792	0.0046	0.0027	0.0507	0.0563
	θ	0.0331	0.1197	0.3201	0.3462	0.0048	0.1223	0.3190	0.3466
1PNO	β	−0.0049	0.0045	0.0658	0.0739	0.0110	0.0024	0.0471	0.0524
	θ	−0.0049	0.1036	0.2976	0.3209	0.0117	0.1029	0.2981	0.3198
CLLM	β	−0.0118	0 .0047	0.0657	0.0726	−0.0078	0.0023	0.0456	0.0516
	θ	−0.0104	0.0943	0.2828	0.3054	−0.0059	0.0954	0.2843	0.3057

The terms Bias, MSE, SD, and SE denote the average bias, mean square error (MSE), standard deviation (SD), and standard error (SE) of the parameters, respectively.

3.4. Bayesian model assessment

The following four model selection criteria will be used in this paper to evaluate the goodness of model fit: (1) DIC, (2) Logarithm of the pseudomarginal Likelihood (LPML; Geisser and Eddy, 1979; Ibrahim et al., 2001), (3) Widely applicable information criterion (WAIC; Watanabe and Opper, 2010), and (4) Leave-one-out cross-validation (LOO; Vehtari et al., 2017). In addition, the last two information criteria are calculated based on the R package **loo** (Vehtari et al., 2017).

4. Simulation studies

4.1. Simulation 1

In this simulation study, our aim is to assess the accuracy of parameter estimation for various one-parameter symmetric and asymmetric IRT models implemented using the Stan software. The following four models will be considered: (1) 1PGlogit($\lambda_{1j}, \lambda_{2j}$), $j = 1, 2, ..., J$; (2) Rasch (1PGlogit(0, 0)); (3) 1PNO (1PGlogit(0.165, 0.165)); and (4) CLLM (1PGlogit(0.62, −0.037)).

4.1.1. Simulation designs

The true values of the parameters are generated following this formulation: $\theta \sim N(0, 1)$, $b \sim N(0, 1)$. For the 1PGlogit($\lambda_{1j}, \lambda_{2j}$) model, the true values of ($\lambda_{1j}, \lambda_{2j}$) are generated from the distribution $\lambda_{1j} \sim N(0, 0.5^2)\mathcal{I}(-1, +\infty)$, $\lambda_{2j} \sim N(0, 0.5^2)\mathcal{I}(-1, +\infty)$. Meanwhile, λ_{1j} is fixed at 0, 0.165, and 0.62 for the Rasch, 1PNO, and CLLM models, respectively, while λ_{2j} is fixed at 0, 0.165, and –0.037, respectively. The manipulated factors include sample size (i.e., the number of students) $N = 1,000$, 2,000, and item length $J = 20$, 40. Thus, there are four simulation conditions for each model, and each simulation condition was replicated 50 times. We set four chains in each simulation, each executing 3,000 iterations, and the burn-in period is 2,000 iterations.

4.1.2. Convergence diagnosis

Firstly, we examined the convergence of the MCMC procedure implemented in **rstan**. As an example, we considered the case with $N = 1,000$ and $J = 20$. The potential scale reduction factor (PSRF; also known as $\hat{R}$, Brooks and Gelman, 1998) values of the parameters in each model are shown in Figure 3, which presents a boxplot of the $\hat{R}$ values for all difficulty parameters

TABLE 2 Comparing the DIC, LPML, WAIC, and LOO values for 1PGlogit, Rasch, 1PNO, and CLLM models in simulation 2.

Fitted model	DIC	LPML	WAIC	LOO
True model: 1PGlogit				
1PGlogit	**23306.55**	**−11686.98**	**23335.52**	**23367.22**
Rasch	23559.69	−11781.79	23560.16	23566.15
1PNO	23546.30	−11790.91	23574.84	23584.28
CLLM	23476.40	−11767.11	23520.99	23535.90
True model: Rasch				
1PGlogit	20965.74	−10489.26	**20974.06**	20980.80
Rasch	20970.16	**−10488.71**	20974.30	**20979.91**
1PNO	**20964.79**	−10500.66	20994.16	21003.703
CLLM	21043.82	−10575.56	21123.53	21150.62
True model: 1PNO				
1PGlogit	22122.17	−10618.52	21226.77	21238.39
Rasch	21250.88	−10622.21	21241.22	21246.88
1PNO	**21201.76**	**−10611.60**	**21215.99**	**21225.56**
CLLM	21278.45	−10684.17	21342.14	21367.91
True model: CLLM				
1PGlogit	22131.63	−11091.62	22148.53	22178.66
Rasch	22330.45	−11155.46	22307.58	22313.39
1PNO	22252.20	−11129.47	22251.93	22261.30
CLLM	**22111.62**	**−11073.75**	**22127.46**	**22147.63**

The bold values represent the minimum values of the corresponding model selection criteria across all candidate models.

across 50 repeated simulations. It can be observed that the $\hat{R}$ for all parameters in each model is close to 1 and less than 1.05, indicating that all parameters have converged. In addition, we selected the parameters for the first item, namely β_1, λ_{11}, λ_{21}, as well as the latent trait of the first student θ_1 and the standard deviations $\sigma_\beta, \sigma_\lambda$. We plotted the MCMC traces of these parameters across the four chains in Figure 4. The red vertical line represents the burn-in value and the colored circles represent the initial values. From the trace plots, it is apparent that all parameters reached stationarity before the burn-in period, which further validates that the convergence is assured when using the Stan software for parameter estimation.

4.1.3. Analysis of parameter estimation accuracy

In this study, we examine the accuracy of the estimation for the item parameters and latent trait parameters of each model. We computed the average bias, MSE, SE, and SD for each parameter, which are presented in Table 1. By examining the results in the table, we draw the following conclusions: First, the estimation appears unbiased, as reflected by the minimal and close-to-zero bias of all parameters. Second, our estimation exhibits large sample properties, meaning the precision of parameter estimation improves as the number of students increases for item parameters, and as the number of items increases for ability parameters. For instance, in the 1PGlogit model, as the sample size increases from N=1,000 to N=2,000, the MSE, SE, and SD of item parameters $\boldsymbol{\beta}$, $\boldsymbol{\lambda_1}$, $\boldsymbol{\lambda_2}$ decrease. Similarly, when increasing from J=20 to J=40, the MSE, SE, and SD of $\boldsymbol{\theta}$ decrease as well. Similar conclusions hold true in the Rasch, 1PNO, and CLLM models. Moreover, we observed that the estimation precision of latent trait parameters $\boldsymbol{\theta}$ is not as robust as that of difficulty parameters $\boldsymbol{\beta}$ across all models. This can be attributed to the limited number of items (only 20 or 40 items). Specifically, in the 1PGlogit model, the estimation precision of $\boldsymbol{\lambda}$ is also poorer than that of $\boldsymbol{\beta}$, and we speculate that this may be due to the interaction between $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ affecting the estimation precision.

4.2. Simulation 2

In this simulation study, our aim is to assess the model fit of traditional symmetric IRT models, asymmetric IRT model, and the Glogit IRT models under the framework of the one-parameter IRT.

We consider a sample size of $N = 1,000$ individuals, with the test length fixed at 20. Item responses are generated within the framework of a one-parameter IRT model. We consider four item response models: (1) 1PGlogit($\lambda_{1j}, \lambda_{2j}$), $j = 1, 2, ..., J$; (2) Rasch (1PGlogit(0, 0)); (3) 1PNO (1PGlogit(0.165, 0.165)); and (4) CLLM (1PGlogit(0.62, −0.037)). Therefore, we evaluate the model fitting in the following four cases.

- Case 1: True model: 1PGlogit($\lambda_{1j}, \lambda_{2j}$) v.s. Fitted model: 1PGlogit($\lambda_{1j}, \lambda_{2j}$), Rasch, 1PNO, and CLLM;
- Case 2: True model: Rasch v.s. Fitted model: 1PGlogit($\lambda_{1j}, \lambda_{2j}$), Rasch, 1PNO, and CLLM;
- Case 3: True model: 1PNO v.s. Fitted model: 1PGlogit($\lambda_{1j}, \lambda_{2j}$), Rasch, 1PNO, and CLLM;
- Case 4: True model: CLMM v.s. Fitted model: 1PGlogit($\lambda_{1j}, \lambda_{2j}$), Rasch, 1PNO, and CLLM.

The true values and prior distributions for the parameters are specified in the same way as in simulation 1. To implement the MCMC sampling algorithm, chains of length 3,000 are chosen, with an initial burn-in period of 2,000. The results of the Bayesian model assessment, based on 50 replications, are shown in Table 2. It is worth noting that the reported results of DIC, LPML, WAIC, and LOO are based on the average of these 50 replications. The corresponding boxplots of the four Bayesian model assessment indexes is shown in Figure 5. Additionally, we have compiled the number of times each model was selected as the best or second-best model in Table 3.

According to Tables 2, 3, when the true model is a 1PGlogit model, the 1PGlogit model is consistently chosen as the optimal model for data fitting based on the average values of the four model evaluation criteria, compared to the other three competing models. The second-best model is mostly the asymmetric CLLM, except for two instances where the Rasch model is selected for LPML and LOO criteria. When the true model is the CLLM model, the evaluation results are very similar to the case where the true model is the 1PGlogit model. With only a few exceptions, the CLLM model is chosen as the optimal model for almost all evaluation indicators, and the 1PGlogit model is chosen as the second-best model. Additionally, from Table 2 and Figure 5,

10.3389/fpsyg.2023.1248454

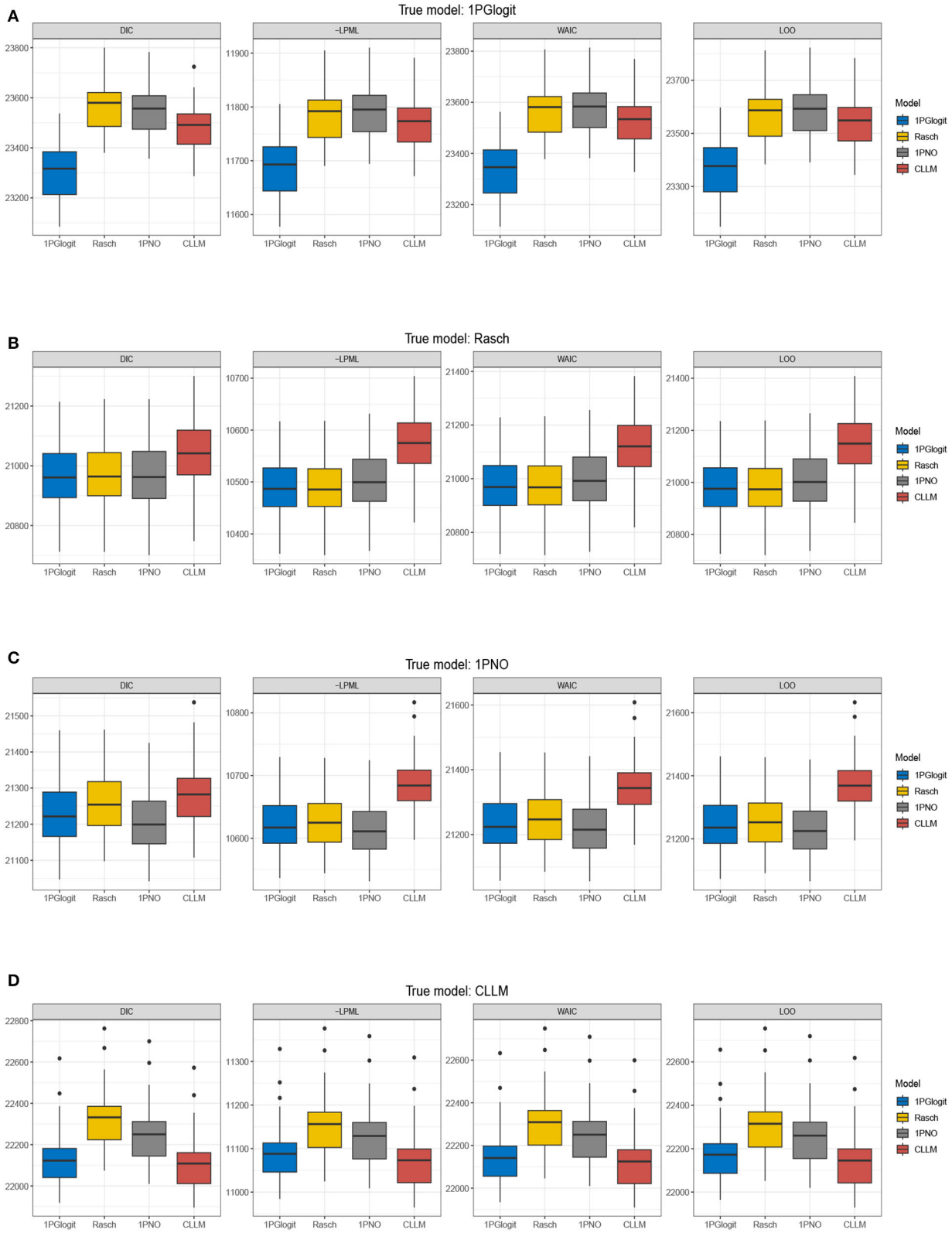


FIGURE 5
Boxplots of DIC, −LPML, WAIC, and LOO for 1PGlogit, Rasch, 1PNO, and CLLM models in simulation 2. **(A)** True model: 1PGlogit model. **(B)** True model: Rasch model. **(C)** True model: 1PNO model. **(D)** True model: CLLM.

we can observe that the fitting results of the 1PGlogit model are not significantly different from that of the CLLM model. In fact, 1PGlogit model has been selected four times as the best model using DIC and WAIC. However, the fitting results of the other two symmetric models, Rasch and 1PNO, are noticeably worse compared with that of the CLLM and 1PGlogit models.

10.3389/fpsyg.2023.1248454

TABLE 3 Number of times selected as the best model and the second-best model based on DIC, LPML, WAIC, and LOO in simulation 2.

	Times of selected as the best model				Times of selected as the second-best model			
	1PGlogit	Rasch	1PNO	CLLM	1PGlogit	Rasch	1PNO	CLLM
True model: 1PGlogit								
DIC	50	0	0	0	0	0	0	50
LPML	50	0	0	0	0	2	0	48
WAIC	50	0	0	0	0	0	0	50
LOO	50	0	0	0	0	2	0	48
True model: Rasch								
DIC	20	2	28	0	25	21	4	0
LPML	13	36	1	0	36	14	0	0
WAIC	24	25	1	0	25	24	1	0
LOO	15	34	1	0	34	16	0	0
True model: 1PNO								
DIC	2	0	48	0	47	2	1	0
LPML	2	0	48	0	37	12	1	0
WAIC	3	0	47	0	42	5	3	0
LOO	3	0	47	0	37	1	1	2
True model: CLLM								
DIC	4	0	0	46	46	0	0	4
LPML	0	0	0	50	0	0	0	50
WAIC	4	0	0	46	46	0	0	4
LOO	0	0	0	50	0	0	0	50

Interestingly, when the true model is the Rasch model, we observe that the fitting results of the 1PGlogit and 1PNO models are highly similar to those of the Rasch model. In terms of average DIC value, the 1PGlogit and 1PNO models even perform better and are often chosen as the best models. The Rasch model has only a very slight advantage over the 1PGlogit model in LPML and LOO, and in many cases, the 1PGlogit model is selected as the true model. The difference between 1PGlogit model and Rasch model, based on the four model evaluation criteria, is very small and less than 1. The fitting results of the 1PNO model are slightly worse than that of 1PGlogit and Rasch models based on LPML, WAIC, and LOO criteria, and the performance of the CLLM is the worst in all four evaluation criteria. In the case where the 1PNO model is the true model, we also observe that the performance of the CLLM is consistently the worst. While the 1PNO model slightly outperforms the 1PGlogit model across all model evaluation criteria, the 1PGlogit model still provides a good fit and has been selected as the best fitting model several times based on these model evaluation criteria.

Additionally, we chose the first item from four simulation conditions, respectively, and plotted their true ICCs against the four fitted ICCs for comparison in Figure 6. The true ICC is represented by the black line, while the red line illustrates the ICC fitted using 1PGlogit model. It can be noted that regardless of the true model type, our 1PGlogit model can provide an excellent fit, especially when the Rasch model and 1PNO model serve as the true model, the ICC fitted by 1PGlogit model almost coincides with the true ICC curve. In summary, 1PGlogit model proves to be a versatile generalized model that fits several widely used one-parameter IRT models effectively.

4.3. Simulation 3

In our previous discussion, we noted that the two shape parameters in the proposed 1PGlogit model can control whether the ICC has a heavy or light tail, playing a role similar to the lower asymptote parameter in the three-parameter IRT models, and the upper asymptote parameter in the more generalized four-parameter IRT models. In this simulation study, we focus on comparing the fit superiority of the 1PGlogit model with the traditional 3PL and 4PL models.

We consider a sample size of $N = 1,000$ individuals, with the test length fixed at 20. Item responses are generated from the 3PL model and 4PL model. Therefore, we evaluate the model fitting in the following two cases.

- Case 1: True model: 3PL v.s. Fitted model: 1PGlogit($\lambda_{1j}, \lambda_{2j}$), Rasch, 1PNO, CLLM, and 3PL;

10.3389/fpsyg.2023.1248454

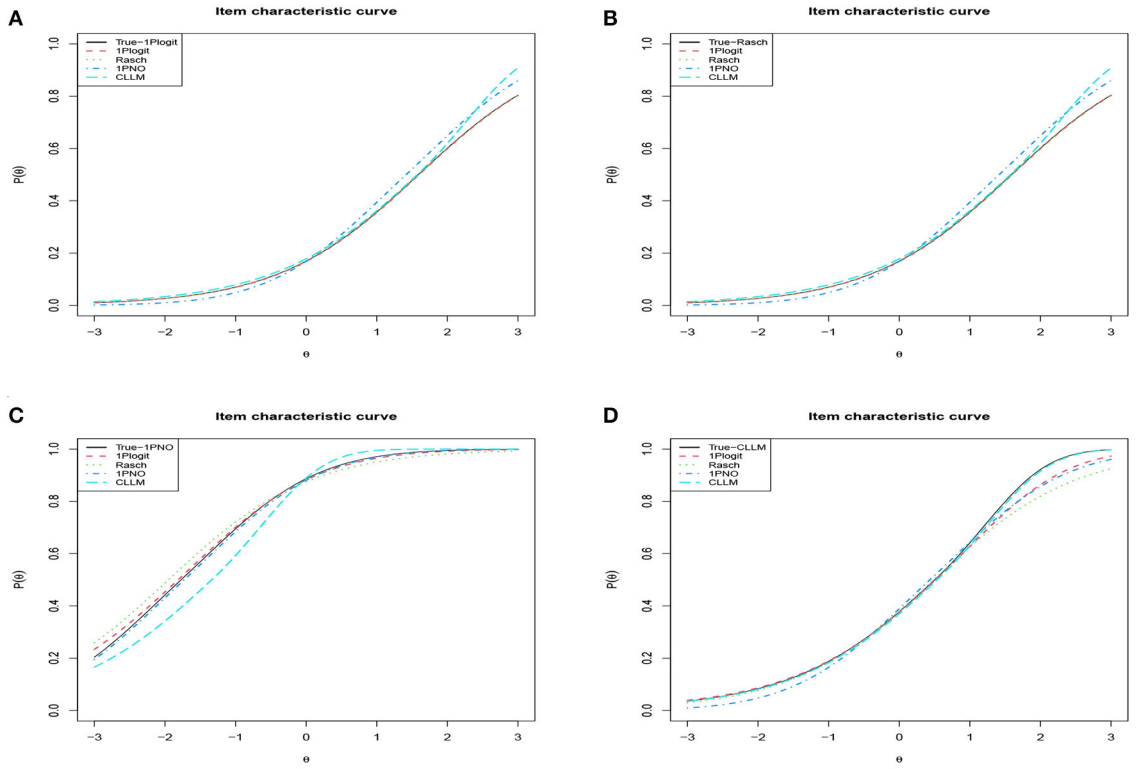


FIGURE 6
Analyzing the degree of fit for ICCs across different true models and fitting models in simulation 2. **(A)** True model: 1PGlogit with $\beta_1 = -0.5734$, $\lambda_{11} = 1$, $\lambda_{21} = 1$. **(B)** True model: Rasch (1PGlogit with $\beta_1 = 1.5891$, $\lambda_{11} = 0$, $\lambda_{21} = 0$). **(C)** True model: 1PNO (1PGlogit with $\beta_1 = -1.7712$, $\lambda_{11} = 0.165$, $\lambda_{21} = 0.165$). **(D)** True model: CLLM (1PGlogit with $\beta_1 = 0.5036$, $\lambda_{11} = 0.62$, $\lambda_{21} = -0.037$).

TABLE 4 Comparing the DIC, LPML, WAIC, and LOO values for 1PGlogit, Rasch, 1PNO, CLLM, 3PL, and 4PL models in simulation 3.

Fitted model	DIC	LPML	WAIC	LOO
True model: 3PL model				
1PGlogit	22205.12	−11136.05	22234.30	22265.62
Rasch	22504.53	−11268.28	22532.98	22533.96
1PNO	22486.74	−11268.22	22528.35	22538.81
CLLM	22469.30	−11290.41	22550.59	22579.95
3PL	**22186.26**	**−11095.16**	**22169.37**	**22189.71**
True model: 4PL model				
1PGlogit	25991.13	−13031.57	26054.59	26062.20
Rasch	26234.96	−13144.81	26286.36	26292.44
1PNO	26257.85	−13164.60	26323.77	26332.01
CLLM	26266.45	−13183.73	26357.70	26369.97
4PL	**25985.41**	**−12933.86**	**25857.94**	**25866.80**

The bold values represent the minimum values of the corresponding model selection criteria across all candidate models.

- Case 2: True model: 4PL v.s. Fitted model: 1PGlogit($\lambda_{1j}, \lambda_{2j}$), Rasch, 1PNO, and CLLM, and 4PL.

The true values of parameters in the 3PL and 4PL models are generated as follows: $\alpha_j \sim \text{U}(0.5, 2)$, $\beta_j \sim \text{N}(0, 1)$, $c_j \sim \text{Beta}(5, 17)$ and $d_j \sim \text{Beta}(17, 5)$ ($d_j = 1$ for 3PL model). The prior distribution of parameters in the 1PGlogit model, Rasch model, 1PNO model, and CLLM are generated the same as in simulation 1. Moreover, we wish to clarify the prior distributions setting for the parameters in the 3PL/4PL models: $\log \alpha_j \sim \text{N}(0, 1)$, $\beta_j \sim \text{N}(0, \sigma_\beta^2)$, $c_j \sim \text{U}(0, 0.5)$, $d_j \sim \text{U}(0, 0.5)$ (in 4PL model), and $\sigma_\beta \sim$ Cauchy$(0, 5)$. To implement the MCMC sampling algorithm, chains of length 5,000 are chosen, with an initial burn-in period of 4,000.

In Table 4, we present the DIC, LPML, WAIC, and LOO values for each model. Figure 7 depicts the boxplots of these four model selection criteria across 50 replications. Additionally, Table 5 summarizes the instances where each model was selected as the best or second best fitting model across the 50 replications. The results indicate that when the true model is the 3PL model, the average values of −LPML, WAIC, and LOO for the 3PL model are the lowest among all models under consideration. In all 50 replications, these evaluation criteria identify the true 3PL model as the best model. For the second-best model selection, apart from LOO (which chose the Rasch model once), all other criteria consistently select the 1PGlogit model. Although the average DIC value for the 3PL model is the lowest, it differs from the other three criteria. In 12 out of 50 replications, the 1PGlogit

10.3389/fpsyg.2023.1248454

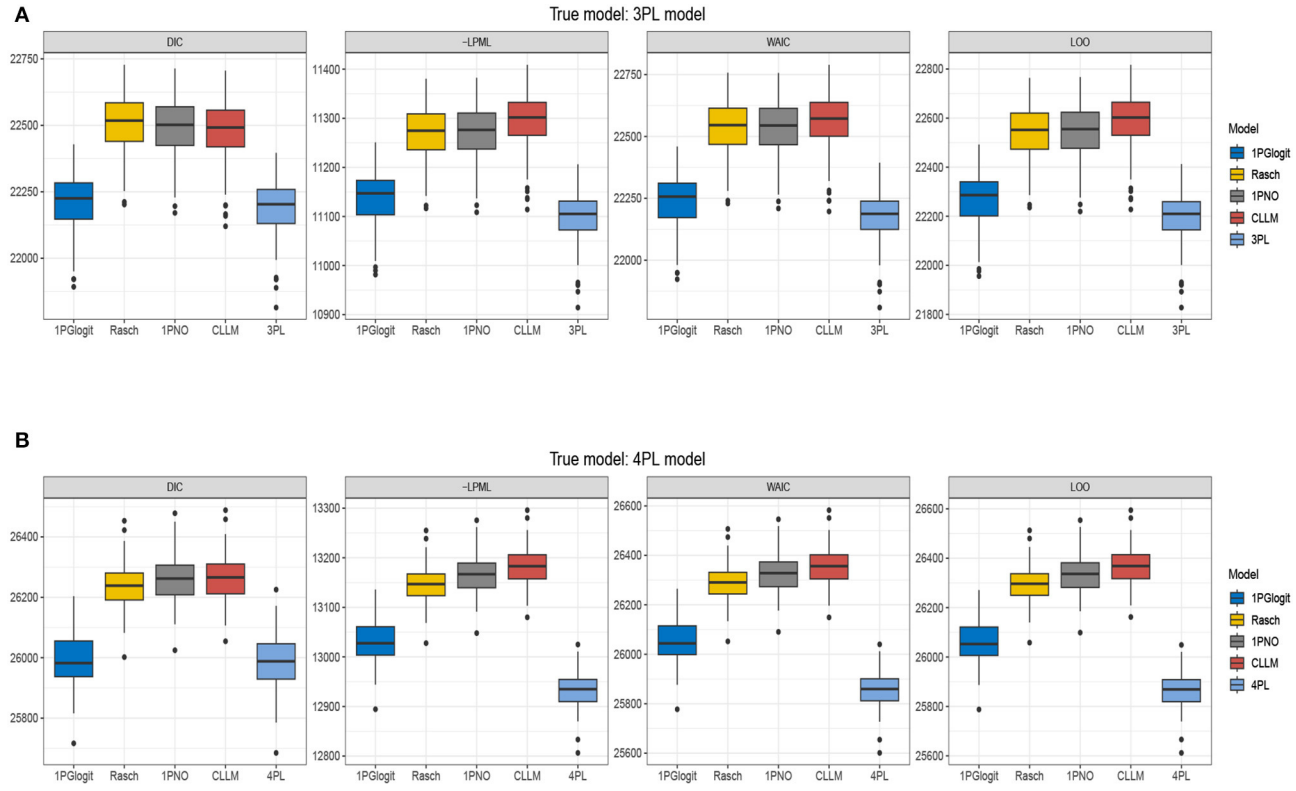


FIGURE 7
Boxplots of DIC, −LPML, WAIC, and LOO for 1PGlogit, Rasch, 1PNO, CLLM, 3PL, and 4PL models in simulation 3. **(A)** True model: 3PL model. **(B)** True model: 4PL model.

TABLE 5 Number of selected times as the best-model and the second-best model based on DIC, LPML, WAIC, and LOO in Simulation 3.

	Times of selected as the best model					Times of selected as the second best model				
	1PGlogit	Rasch	1PNO	CLLM	3PL	1PGlogit	Rasch	1PNO	CLLM	3PL
True model: 3PL model										
DIC	12	0	0	0	38	38	0	0	0	12
LPML	0	0	0	0	50	50	0	0	0	0
WAIC	0	0	0	0	50	50	0	0	0	0
LOO	0	0	0	0	50	49	1	0	0	0
	1PGlogit	Rasch	1PNO	CLLM	3PL	1PGlogit	Rasch	1PNO	CLLM	4PL
True model: 4PL model										
DIC	25	0	0	0	25	25		0	0	25
LPML	0	0	0	0	50	50	0	0	0	0
WAIC	0	0	0	0	50	50	0	0	0	0
LOO	0	0	0	0	50	50	0	0	0	0

model is selected as the best model, and in 38 replications, it's chosen as the second-best model. These findings suggest that our flexible 1PGlogit model can effectively fit the 3PL model. Considering the values of various model selection criteria and the boxplot results, the fitting performance of the 1PGlogit model is significantly superior to other one-parameter models. To further illustrate this, we plotted the ICC of the first item for the true 3PL model, as well as ICC curves fitted by the five different models in Figure 8. The plots reveal that, aside from the fitted 3PL model, our 1PGlogit model shows the best fit with the true ICC, regardless of item difficulty. In the 3PL model, the assumed guessing behavior causes the lower asymptote of its ICC to be above zero. Our 1PGlogit model can account for this phenomenon through the parameter λ, suggesting that our model can also interpret the assumed guessing behavior inherent in the 3PL model.

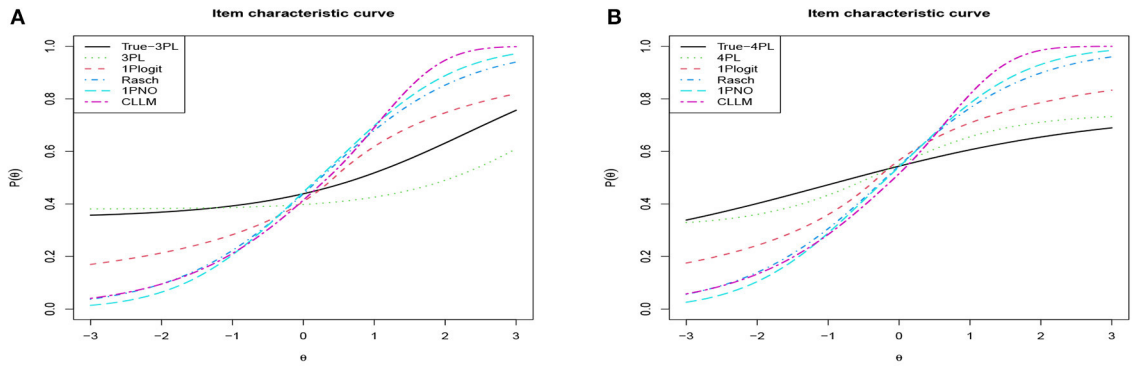


FIGURE 8
Analyzing the degree of fit for ICCs across different true models and fitting models in simulation 3. **(A)** True model: 3PL model with $\alpha_1 = 0.7766$, $\beta_1 = 2.3315$, $c_1 = 0.3470$. **(B)** True model: 4PL model with $\alpha_1 = 0.5167$, $\beta_1 = -1.0322$, $c_1 = 0.1894$, $d_1 = 0.7518$.

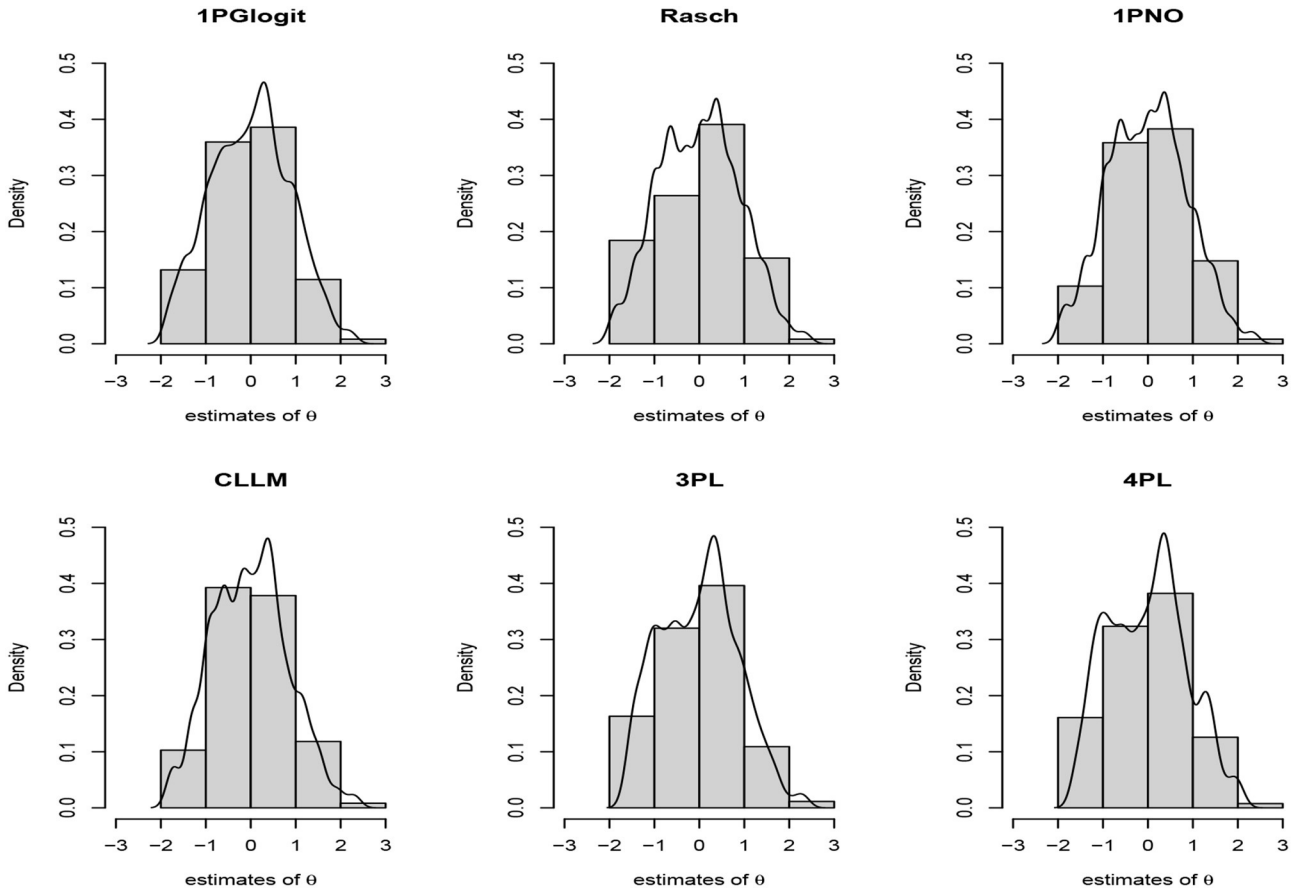


FIGURE 9
Item characteristic curve (ICC) of all items based on 1PGlogit model for the real data.

Secondly, when the true model is the 4PL model, the results are nearly identical to those under the 3PL model. The 4PL model performs the best based on LPML, WAIC, and LOO, and is selected as the optimal model in all 50 repetitions. The second-best model is consistently the 1PGlogit model. In terms of DIC value, the average for the 4PL model is the lowest, but in 25 out of 50 repetitions, the 1PGlogit model is chosen as the best. As illustrated by the boxplot in Figure 7, the model selection criteria of the 1PGlogit model are significantly lower than those of the other one-parameter models. Figure 8 displays the ICCs of the first item. Aside from the 4PL model, the ICC of the 1PGlogit model demonstrates the best fitting performance, suggesting that this flexible 1PGlogit model provides a well-fitted representation of the guessing behavior and slipping

TABLE 6 Values of DIC, LPML, WAIC, and LOO for 1PGlogit, Rasch, 1PNO, CLLM, 3PL, and 4PL models for the real data.

Model	DIC	LPML	WAIC	LOO
1PGlogit	**6464.903**	**−3257.137**	**6454.856**	**6499.141**
Rasch	6689.912	−3341.841	6679.935	6684.846
1PNO	6647.231	−3330.087	6651.59	6661.054
CLLM	6708.316	−3387.79	6742.526	6771.13
3PL	6528.033	−3276.638	6528.809	6552.125
4PL	6552.333	−3275.981	6528.991	6550.834

The bold values represent the minimum values of the corresponding model selection criteria across all candidate models.

behavior assumed in the 4PL model, which affects the lower and upper asymptotes.

In summary, the 1PGlogit model demonstrates superior fitting performance for asymmetric models compared to other one-parameter models. This model enhances flexibility by adjusting the parameter $\boldsymbol{\lambda}$ to fit the upper and lower asymptotes. However, we observed that DIC sometimes failed to identify the true model in this simulation, as was also the case when Rasch was the true model in Simulation 2. According to Luo and Al-Harbi (2016), within the dichotomous IRT framework, the performances of WAIC and LOO surpass that of DIC. Therefore, in light of the findings of this paper, we recommend giving greater consideration to LPML, WAIC, and LOO criteria when selecting models.

5. Real data

For this example, we use the 2015 computer-based PISA science data. Out of all the countries that took part in the computer-based science assessment, we selected data from the United States of America (USA). The initial sample consisted of 685 students, but 76 students were excluded due to Not Reached (original code 6) or No Response (original code 9) outcomes. These Not Reached and No Response results were treated as missing data. Therefore, the final sample size stands at 609 students, for whom the response data is available. The 11 items were scored on a dichotomous scale. We utilize six different models to fit the PISA dataset. This includes two symmetric models, namely the Rasch and the 1PNO models, in conjunction with three asymmetric models: the CLLM, the 3PL model, the 4PL model, and our generalized logistic model, known as the 1PGlogit model. During the process of estimation, we employ the same prior probabilities for the unknown parameters as used in simulations 2 and 3. Throughout all Bayesian computations, we generate 5,000 MCMC samples after a burn-in period of 4,000 iterations for each model to compute all the posterior estimates. The convergence of the chains is assured by evaluating the PSRF values ($\hat{R}$). For each model, the PSRF values of all parameters, both item and person, are observed to be under 1.1.

First, we depicted the frequency distribution histogram of the estimated ability parameter $\boldsymbol{\theta}$ values across different models in Figure 9, and fitted their respective distribution curves. From this, it is apparent that the distributions of the estimated ability parameters remain largely consistent across the varied models.

TABLE 7 Parameter estimates for all items based on the 1PGlogit model in real data.

	Estimate	SD	HPDI	$\hat{R}$
$\boldsymbol{\beta}$				
Item 1	−0.3259	0.1178	[−0.5528, −0.1009]	1.0027
Item 2	0.7981	0.1544	[0.5035, 1.1023]	1.0001
Item 3	0.6522	0.1241	[0.4250, 0.9101]	1.0024
Item 4	−0.1680	0.1069	[−0.3772, 0.0352]	1.0006
Item 5	−0.7112	0.1055	[−0.9126, −0.4937]	1.0015
Item 6	2.4805	0.2461	[2.0107, 2.9589]	0.9996
Item 7	0.0470	0.1325	[−0.2003, 0.3080]	1.0000
Item 8	−0.5728	0.1416	[−0.8627, −0.3006]	0.9999
Item 9	0.9687	0.1407	[0.7115, 1.2589]	1.0010
Item 10	1.5533	0.1419	[1.2907, 1.8341]	1.0010
Item 11	−2.5073	0.3414	[−3.2047, −1.8844]	1.0001
$\boldsymbol{\lambda_1}$				
Item 1	0.4927	0.2843	[−0.0540, 1.0522]	1.0018
Item 2	−0.2222	0.4696	[−0.9962, 0.6246]	1.0018
Item 3	−0.1674	0.4650	[−0.9995, 0.6549]	1.0023
Item 4	0.4861	0.3014	[−0.0436, 1.1152]	1.0006
Item 5	1.2090	0.3863	[0.5059, 1.9562]	1.0010
Item 6	0.1382	0.6449	[−0.9873, 1.3015]	1.0008
Item 7	−0.0362	0.3830	[−0.8312, 0.7113]	1.0020
Item 8	−0.0904	0.2658	[−0.6508, 0.4159]	1.0014
Item 9	−0.0398	0.5547	[−0.9744, 0.9865]	1.0017
Item 10	0.2537	0.6769	[−0.9908, 1.4677]	1.0000
Item 11	−0.4990	0.1685	[−0.8577, −0.1966]	1.0016
$\boldsymbol{\lambda_2}$				
Item 1	0.4891	0.5216	[−0.5493, 1.5489]	1.0029
Item 2	−0.3930	0.2545	[−0.9086, 0.0654]	1.0020
Item 3	0.9644	0.5037	[0.0353, 1.9567]	1.0046
Item 4	0.7601	0.3364	[0.1513, 1.4065]	1.0017
Item 5	0.6816	0.6276	[−0.4691, 1.9689]	1.0003
Item 6	0.7804	0.3510	[0.2051, 1.4938]	0.9994
Item 7	−0.2759	0.3416	[−0.9636, 0.3079]	0.9997
Item 8	0.2539	0.5746	[−0.9122, 1.2491]	1.0002
Item 9	0.3990	0.2278	[−0.0012, 0.8625]	1.0019
Item 10	1.1567	0.0094	[0.4972, 1.9112]	1.0010
Item 11	0.1632	0.6592	[−0.9988, 1.3891]	1.0000

Upon examining the fitted distributions of the estimated $\boldsymbol{\theta}$, it can be observed that the ability distribution under the 1PGlogit model is closest to a normal distribution. The $\boldsymbol{\theta}$ distributions under the Rasch and 1PNO models are notably similar, while the $\boldsymbol{\theta}$ distributions under the 3PL model are more analogous to those of the 4PL model. Next, we provide detailed results of the Bayesian

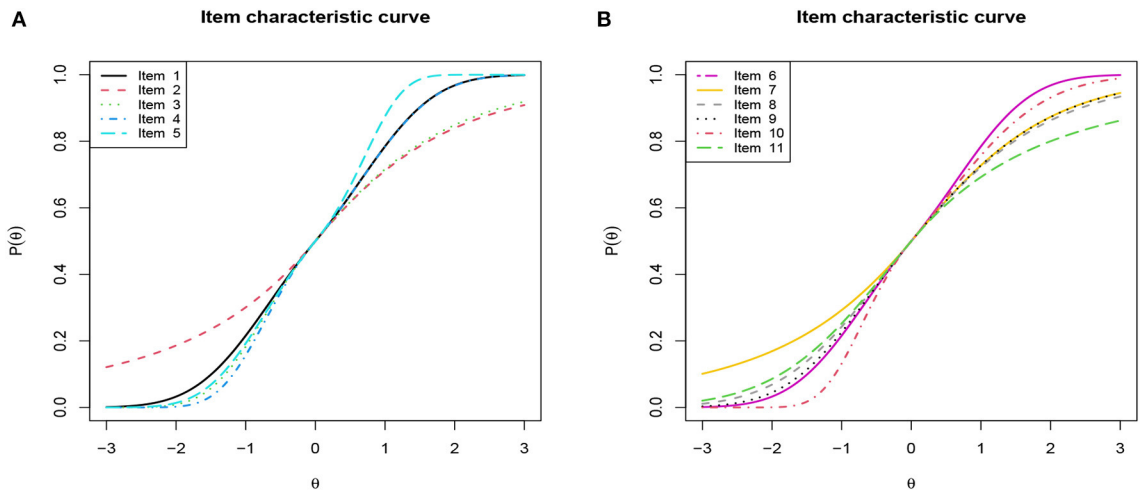


FIGURE 10
Item characteristic curve (ICC) of all items based on 1PGlogit model for the real data. **(A)** ICC of Item 1–5. **(B)** ICC of Item 6–11.

model assessment for the PISA dataset in Table 6. All these criteria indicate that the 1PGlogit model fits the data best among the six models. The second-best fitting model tends to be either the 3PL model or the 4PL model, both of which demonstrate similar fitting effects, while the three one-parameter IRT models show a notably inferior fit compared to the others. Hence, we surmise that the data shows a preference for flexible asymmetrical models. Based on the results of the model assessment, we will proceed with the best fitting 1PGlogit model for the analysis of the PISA data. In Table 7, we provide the estimated values of parameters in the 1PGlogit model, including the SD, 95% highest posterior density interval (HPDI), and $\hat{R}$ for each parameter. It is evident from the $\hat{R}$ values that the Markov chain has achieved convergence. Examining the estimated parameter values, we note firstly that item 6 is the most difficult, with $\beta_6 = 2.4805$, while item 11 is the easiest, with $\beta_{11} = -2.5073$. Moreover, for the parameter $\boldsymbol{\lambda_1}$, the values are mostly small, except for item 5 which exceeds 1, suggesting that the tail of this item's ICC approaches the upper asymptote more quickly. Conversely, the estimated values for $\boldsymbol{\lambda_2}$ are generally larger and positive, such as for item 10, which exceeds 1, indicating a rapid approach to the lower asymptote for the tail of its ICC. Lastly, we have plotted the ICCs for all the items in Figure 10. From Figure 10, it can be seen that for item 2, there appears to be some guessing behavior among low ability students, as they have a certain probability of answering the item correctly even with very low ability. Conversely, high ability students may exhibit slipping behavior, as even with relatively high ability, their probability of answering correctly is only around 90%. In contrast, for item 5, students with ability values below 2 have virtually no chance of answering correctly, while those with ability values exceeding 1.5 have almost no chance of answering incorrectly. In essence, the 1PGlogit model can deliver robust data fitting and outstanding interpretability.

6. Discussion

This paper discusses a generalized one-parameter IRT model, the 1PGlogit model, which can encompass commonly-used IRT models such as the Rasch, 1PNO, and the recently proposed CLLM as its submodels. Owing to its adjustable parameter $\boldsymbol{\lambda}$, it exhibits high flexibility, which enables control over the rate at which it approaches the upper and lower asymptotes of the ICC. In this paper, we first examine the accuracy of the model in parameter recovery using the Stan program. Subsequently, we investigate its performance in fitting data generated by other one-parameter IRT models. Finally, we delve deeper into its effectiveness in fitting asymmetric 3PL and 4PL models.

From the simulation results, we can draw the following conclusions. Firstly, the estimates generated by Stan are consistent with the large sample properties and exhibit excellent parameter recovery accuracy. The difficulty parameter demonstrates the highest estimation precision, followed by $\boldsymbol{\lambda}$ and θ. Secondly, the 1PGlogit model showcases commendable fitting performance for data generated by its various submodels. It ranks as the best model in terms of fitting performance, with the exception of the true model. Finally, the 1PGlogit model presents an outstanding fit for data generated by the asymmetric 3PL and 4PL models, markedly superior to other one-parameter IRT models. The 1PGlogit model can more accurately recover the shape of the ICC of the 3PL/4PL model.

In summary, the 1PGlogit model is a highly flexible and generalized model that encompasses Rasch, 1PNO, and CLLM as its submodels. Its parameter $\boldsymbol{\lambda}$ adjusts the speed at which the ICC curve approaches the upper and lower asymptotes. A larger λ_1 results in a quicker approach to the upper asymptote, and a larger λ_2 results in a swifter approach to the lower asymptote. As such, the 1PGlogit model can effectively accommodate the assumptions of guessing and slipping behavior in the 3PL and 4PL models, which would otherwise cause the upper and lower asymptotes to diverge from 1 and 0, respectively. However, the 1PGlogit model also has its limitations. Firstly, the constraint that its parameter $\boldsymbol{\lambda}$ must be greater than -1 may inhibit the model's ability to depict behaviors on the ICC where the asymptotes significantly diverge from 1 and 0. Secondly, although the 1PGlogit model is a generalized model that includes other one-parameter IRT models, the introduction of the new parameter $\boldsymbol{\lambda}$ adds complexity to the model, and the

estimation accuracy of 1PGlogit is slightly lower than that of other one-parameter models. Moreover, the introduction of $\boldsymbol{\lambda}$ may also introduce some identifiability issues to the model, where $\boldsymbol{\lambda}$ and θ might mutually influence each other.

In conclusion, we would like to propose some directions for future work. The 1PGlogit model is a flexible and generalized model, and this paper merely provides an initial exploration of its advantages in fitting various types of data. We believe there is significant potential for its further development and application, such as extending the 1PGlogit model to higher-order IRT models, graded response models, multilevel IRT models, and longitudinal IRT models, among others. Therefore, in our future research, we will dedicate ourselves to the advancement and application of the 1PGlogit model in these proposed areas. Moreover, a wealth of scholarly work has been dedicated to formulating link functions for binary and ordinal response data. Notable contributions in this field have been made by Aranda-Ordaz (1981), Guerrero and Johnson (1982), Stukel (1988), Kim et al. (2008), Wang and Dey (2010), and Jiang et al. (2014), among others. It is worth exploring whether these existing link functions can be directly applied to the field of IRT. We intend to investigate this possibility in our future work.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.oecd.org/pisa/data/.

Author contributions

XW and JZ completed the writing of this article. XW, JZ, and JL completed the article revisions. JL provided original thoughts. JZ, JL, GC, and NS provided key technical support. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Jilin Province Education Science 14th Five-Year Plan 2022 Annual General Topic + Dynamic Education Quality Monitoring and Evaluation Research Supported by Data - Taking Jilin Province Mathematics Academic Ability Growth Assessment as an Example + Project Approval Number GH22415 and Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515011899).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1248454/full#supplementary-material

References

Aranda-Ordaz, F. J. (1981). On two families of transformations to additivity for binary response data. *Biometrika* 68, 357–363. doi: 10.1093/biomet/68.2.357

Arenson, E. A., and Karabatsos, G. (2018). A Bayesian beta-mixture model for nonparametric IRT (BBM-IRT). *J. Modern Appl. Stat. Methods* 17, 1–18. doi: 10.22237/jmasm/1531318047

Baker, F. B., and Kim, S. H. (2004). *Item Response Theory: Parameter Estimation Techniques, 2nd Edn.* CRC Press. doi: 10.1201/9781482276725

Bechtel, G. G. (1985). Generalizing the Rasch model for consumer rating scales. *Market. Sci.* 4, 62–73. doi: 10.1287/mksc.4.1.62

Bezruczko, N. (2005). *Rasch Measurement in Health Sciences.* Maple Grove, MN: Jam Press.

Bolfarine, H., and Bazan, J. L. (2010). Bayesian estimation of the logistic positive exponent IRT model. *J. Educ. Behav. Stat.* 35, 693–713. doi: 10.3102/1076998610375834

Brooks, S. P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* 7, 434–455. doi: 10.1080/10618600.1998.10474787

Chen, M. H., Dey, D. K., and Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data. *J. Am. Stat. Assoc.* 94, 1172–1186. doi: 10.1080/01621459.1999.10473872

Chen, M. H., Dey, D. K., and Wu, Y. (2002). On robustness of choice of links in binomial regression. *Calcutta Stat. Assoc. Bull.* 53, 145–164. doi: 10.1177/0008068320020113

Duncan, K., and MacEachern, S. (2008). Nonparametric Bayesian modelling for item response. *Stat. Modell.* 8, 41–66. doi: 10.1177/1471082X0700800104

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists.* Lawrence Erlbaum Associates.

Ferguson, G. A. (1942). Item selection by the constant process. *Psychometrika* 7, 19–29. doi: 10.1007/BF02288601

Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection. *J. Am. Statist. Assoc.* 74, 153–160. doi: 10.1080/01621459.1979.10481632

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration. *IEEE Trans. Pattern Anal. Mach. Intell.* 6, 721–741.

Goldstein, H. (1980). Dimensionality and the fitting of unidimensional item response models to multidimensional data. *Appl. Psychol. Meas.* 4, 355–365.

Guerrero, V. M., and Johnson, R. A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika* 69, 309–314. doi: 10.1093/biomet/69.2.309

Hoffman, M. D., and Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* 15, 1593–1623. Available online at: https://jmlr.org/papers/v15/hoffman14a.html

Ibrahim, J. G., Chen, M. H., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York, NY: Springer.

Jiang, X., Dey, D. K., Prunier, R., Wilson, A. M., and Holsinger, K. E. (2014). A new class of flexible link functions with application to species co-occurrence in Cape floristic region. *Ann. Appl. Stat.* 7, 2180–2204 doi: 10.1214/13-AOAS663

Karabatsos, G. (2016). "Bayesian nonparametric IRT," in *Handbook of Item Response Theory,* Vol. 1, ed W. J. van der Linden (Boca Raton, FL: Chapman and Hall/CRC), 323–336.

Kim, S., Chen, M.-H., and Dey, D. K. (2008). Flexible generalized t-link models for binary response data. *Biometrika* 95, 93–106. doi: 10.1093/biomet/asm079

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proc. R. Soc. Edinburgh* 61, 273–287. doi: 10.1017/S0080454100006282

Lawley, D. N. (1944). The factorial invariance of multiple item tests. *Proc. R. Soc. Edinburgh*, 62-A, 74–82. doi: 10.1017/S0080454100006440

Lord, F. M. (1952). A theory of test scores. *Psychometr. Monogr.* 7, 1–100.

Lord, F. M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika* 18, 57–75. doi: 10.1007/BF02289028

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Menlo Park, CA: Addison-Wesley.

Lucke, J. F. (2014). "Positive trait item response models," in *New Developments in Quantitative Psychology: Presentations from the 77th Annual Psychometric Society Meeting*, eds R. E. Millsap, L. A. van der Ark, D. M. Bolt, and C. M. Woods (New York, NY: Routledge), 199–213. doi: 10.1007/978-1-4614-9348-8_13

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* 10, 325–337. doi: 10.1023/A:1008929526011

Luo, Y., and Al-Harbi, K. (2016). "Performances of LOO and WAIC as IRT model selection methods," in *Paper presented at the International Meeting of Psychometric Society* (Ashville, NC).

Luo, Y., and Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educ. Psychol. Meas.* 78, 384–408. doi: 10.1177/0013164417693666

Luzardo, M., and Rodriguez, P. (2015). "A nonparametric estimator of a monotone item characteristic curve," in *Quantitative Psychology Research*, eds L. A. van der Ark, D. Bolt, W. C. Wang, A. Douglas, and S. M. Chow (Cham: Springer International Publishing), 99–108. doi: 10.1007/978-3-319-19977-1_8

Magnus, B. E., and Liu, Y. (2018). A zero-inflated Box-Cox normal unipolar item response model for measuring constructs of psychopathology. *Appl. Psychol. Meas.* 42, 571–589. doi: 10.1177/0146621618758291

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092. doi: 10.1063/1.1699114

Moral, F. J., and Rebollo, F. J. (2017). Characterization of soil fertility using the Rasch model. *J. Soil Sci. Plant Nutr.* doi: 10.4067/S0718-95162017005000035

Mosier, C. L. (1940). Psychophysics and mental test theory: fundamental postulates and elementary theorems. *Psychol. Rev.* 47, 355–366. doi: 10.1037/h0059934

Mosier, C. L. (1941). Psychophysics and mental test theory. II. The constant process. *Psychol. Rev.* 48, 235–249. doi: 10.1037/h0055909

Neal, R. M. (2011). "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, Vol. 2, ed S. Brooks (Boca Raton, FL: CRC Press/Taylorand Francis), 113–162. doi: 10.1201/b10905-6

Plummer, M. (2003). "JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (Vienna), 20–22.

Qin, L. (1998). *Nonparametric Bayesian models for item response data* (Unpublished doctoral dissertation). The Ohio State University, Columbus, OH, United States.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rasch, G. (1960). *Probabilistic Model for Some Intelligence and Achievement Tests.* Copenhagen: Danish Institute for Educational Research.

Richardson, M. W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika* 1, 33–49. doi: 10.1007/BF02288003

Samejima, F. (1997). "Ability estimates that order individuals with consistent philosophies," in *Paper presented at the 1997 Meeting of the American Educational Research Association* (Chicago, IL).

Samejima, F. (1999). "Usefulness of the logistic positive exponent family of models in educational measurement," in *Paper presented at the 1999 Meeting of the American Educational Research Association* (Montreal, QC).

Samejima, F. (2000). Logistic positive exponent family of models: virtue of asymmetric item characteristics curves. *Psychometrika* 65, 319–335. doi: 10.1007/BF02296149

Shim, H., Bonifay, W., and Wiedermann, W. (2022). Parsimonious asymmetric item response theory modeling with the complementary log-log link. *Behav. Res. Methods* 55, 200–219. doi: 10.3758/s13428-022-01824-5

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64, 583–639. doi: 10.1111/1467-9868.00353

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2010). *OpenBUGS Version 3.1.1 User Manual.*

Stan Development Team (2019). *Stan Modeling Language: User's Guide and Reference Manual.*

Stukel, T. A. (1988). Generalized logistic models. *J. Am. Stat. Assoc.* 83, 426–431. doi: 10.1080/01621459.1988.10478613

Tucker, L. R. (1946). Maximum validity of a test with equivalent items. *Psychometrika* 11, 1–13. doi: 10.1007/BF02288894

van der Linden, W. J., and Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York, NY: Springer. doi: 10.1007/978-1-4757-2691-6

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432. doi: 10.1007/s11222-016-9696-4

Wang, X., and Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *Ann. Appl. Stat.* 4, 2000–2023. doi: 10.1214/10-AOAS354

Watanabe, S., and Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 11, 3571–3594. Available online at: https://www.jmlr.org/papers/volume11/watanabe10a/watanabe10a.pdf

Wright, B. D. (1977). Solving measurement problems with the Rasch mode. *J. Educ. Measure.* 14, 97–116. doi: 10.1111/j.1745-3984.1977.tb00031.x

Zhang, J., Zhang, Y. Y., Tao, J., and Chen, M. H. (2022). Bayesian item response theory models with flexible generalized logit links. *Appl. Psychol. Meas.* 46, 382–405. doi: 10.1177/01466216221089343

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 15 September 2023
DOI 10.3389/fpsyg.2023.1216483

OPEN ACCESS

EDITED BY
Ioannis Tsaousis,
National and Kapodistrian University of Athens, Greece

REVIEWED BY
Ulrich S. Tran,
University of Vienna, Austria
James Stamey,
Baylor University, United States

*CORRESPONDENCE
E. F. Haghish
✉ Haghish@uio.no

RECEIVED 03 May 2023
ACCEPTED 24 August 2023
PUBLISHED 15 September 2023

CITATION
Haghish EF, Laeng B and Czajkowski N (2023) Are false positives in suicide classification models a risk group? Evidence for "true alarms" in a population-representative longitudinal study of Norwegian adolescents. *Front. Psychol.* 14:1216483. doi: 10.3389/fpsyg.2023.1216483

Are false positives in suicide classification models a risk group? Evidence for "true alarms" in a population-representative longitudinal study of Norwegian adolescents

E. F. Haghish[1]*, Bruno Laeng[1,2] and Nikolai Czajkowski[1,3]

[1]Faculty of Social Sciences, Department of Psychology, University of Oslo, Oslo, Norway, [2]Faculty of Humanities, RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion, University of Oslo, Oslo, Norway, [3]Division of Mental and Physical Health, Department of Mental Disorders, Norwegian Institute of Public Health, Oslo, Norway

Introduction: False positives in retrospective binary suicide attempt classification models are commonly attributed to sheer classification error. However, when machine learning suicide attempt classification models are trained with a multitude of psycho-socio-environmental factors and achieve high accuracy in suicide risk assessment, false positives may turn out to be at high risk of developing suicidal behavior or attempting suicide in the future. Thus, they may be better viewed as "true alarms," relevant for a suicide prevention program. In this study, using large population-based longitudinal dataset, we examine three hypotheses: (1) false positives, compared to the true negatives, are at higher risk of suicide attempt in future, (2) the suicide attempts risk for the false positives increase as a function of increase in specificity threshold; and (3) as specificity increases, the severity of risk factors between false positives and true positives becomes more similar.

Methods: Utilizing the Gradient Boosting algorithm, we used a sample of 11,369 Norwegian adolescents, assessed at two timepoints (1992 and 1994), to classify suicide attempters at the first time point. We then assessed the relative risk of suicide attempt at the second time point for false positives in comparison to true negatives, and in relation to the level of specificity.

Results: We found that false positives were at significantly higher risk of attempting suicide compared to true negatives. When selecting a higher classification risk threshold by gradually increasing the specificity cutoff from 60% to 97.5%, the relative suicide attempt risk of the false positive group increased, ranging from minimum of 2.96 to 7.22 times. As the risk threshold increased, the severity of various mental health indicators became significantly more comparable between false positives and true positives.

Conclusion: We argue that the performance evaluation of machine learning suicide classification models should take the clinical relevance into account, rather than focusing solely on classification error metrics. As shown here, the so-called false positives represent a truly at-risk group that should be included in suicide prevention programs. Hence, these findings should be taken into consideration when interpreting machine learning suicide classification models as well as planning future suicide prevention interventions for adolescents.

KEYWORDS

suicide attempt, classification error, suicide risk, adolescents, false positive

Introduction

In recent years, the application of supervised machine learning methods has led to a considerable improvement in the accuracy of suicide attempt classification (Franklin et al., 2017; Walsh et al., 2017; Burke et al., 2020; Healy, 2021; Ley et al., 2022; Haghish et al., 2023). Although suicide attempts tend to have a low prevalence in population-representative samples, even a small detection error rate could result in a large number of misclassifications. In particular, a substantial portion of the classification error would constitute False Positives (FP, falsely labeled as suicidal) since most of the population is comprised of non-suicidal people. In recent research on machine learning suicide classification, FP are deemed irrelevant for intervention and are not considered to be at risk of attempting suicide. For example, Linthicum et al. (2019, p. 220) underscored that: "*In the case of false positives, individuals who are not at risk will be classified as being at risk*," a point that is also emphasized in van Vuuren et al. (2021, p. 1418) paper: "*whether it is acceptable to label … [FP] as at risk, when they are actually not.*"

Recently, Haghish and Czajkowski (2023) proposed a theoretical explanation as to why, when numerous psycho-socio-environmental risk factors are incorporated into machine learning retrospective suicide attempt classification models, false positives may be at a higher risk of future suicidal behavior compared to true negatives. As summarized below in Figure 1, they considered three preconditions: (1) causal relationships between predictors and the binary outcome (i.e., suicide attempt) are expected to persistently influence the outcome over time; (2) high accuracy for the model ensuring that the estimated suicide attempt risk is accurate; and (3) a high level of specificity for the classification based on estimated probabilities. Based on these assumptions, Haghish and Czajkowski (2023) postulated that, for accurate models of suicide attempts classification that are trained with a multitude of risk factors and when the specificity threshold of the model is set to be high (i.e., a high cutoff value for classification is considered), it is likely that FP would be at high risk of attempting suicide in the future.

In the present study, we test this hypothesis using a comprehensive population-based longitudinal data from Norwegian adolescents. We develop a model for suicide attempt classification based on the first time point (T1) data and identify FP and TN. Next, we examine the prevalence with which FP and TN report suicide attempts for the first time within a 2-year frame at the second time point (T2) data and compare their respective suicide attempt risks. Specifically, we address three hypotheses: (1) the prevalence of suicide attempts at T2 will be notably higher among FP compared to TN; (2) within the FP group, adolescents with a higher risk score at T1 will more likely report a first-time suicide attempt at T2; (3) The reason we expect such a trend is that, as classification is made based on higher thresholds of estimated suicide attempts risk scores, the severity of known suicide attempt risk factors (e.g., depression, anxiety, non-suicidal self-harm, and suicide ideation) also increases among FP. Thus, an increase in similarity between FP and TP would account for the machine learning model classifying the FP as suicidal (Haghish and Czajkowski, 2023). Simply put, with higher cutoff values, the similarity of FP and TP groups on different risk factors will increase. This understanding of false positives renders the FP a group of interest, both from a methodological and clinical point of view. To our knowledge, this is the first research article to examine whether FP, in the context of machine-learning retrospective suicide classification constitutes a risk group and how this risk is influenced by the choice of specificity.

Methods

Sample

We analyze data from the *Young in Norway* study (https://ung-i-norge.no). In 1992, 11,369 adolescents (5,630 girls and 5,739 boys) from 67 schools in different municipalities participated in the study. A minority of participants did not respond to the suicide attempt item and were dropped from the analyses; i.e. 537 (4.72%, 307 boys and 230 girls). The remaining 10,832 participants ranged in age from 12 to 20 years (mean = 15.75, SD = 1.90). We refer to this sample as T1, by being collected in the first time point. The second wave of data collection was carried out 2 years later in 1994 and 8,018 participants responded to the questionnaire, of which, 593 participants (7.40%) did not answer the suicide attempt item and were excluded from the analyses. We refer to this sample as T2 throughout the article.

Measures

The questionnaire contained items assessing the adolescents' socio-demographic background such as family affluence and cultural capital (Bourdieu, 2018), personal development (e.g., puberty, sexual activities, physical disabilities, etc.), family learning environment (Marjoribanks,

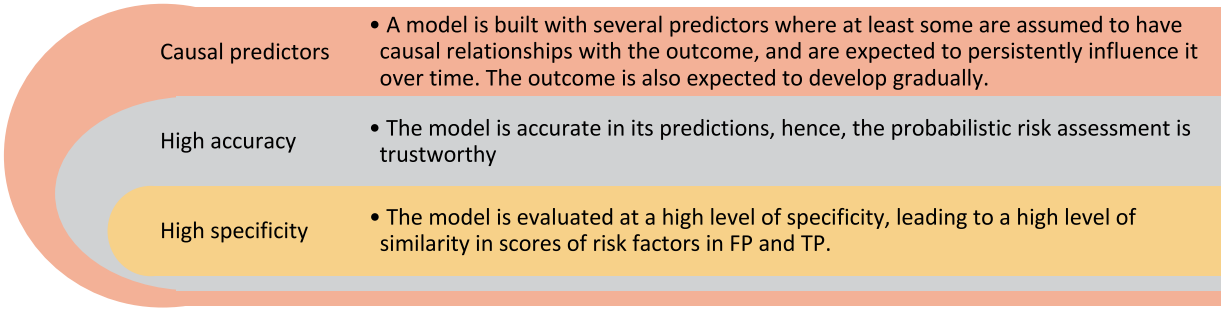


FIGURE 1
Necessary conditions to conceptualize FP as a risk group.

1987), school environment, academic performance, and educational self-efficacy. In addition, adolescents' attitude toward their future occupation was measured with occupational aspiration (Storvoll and Wichstrøm, 2002), career incentives (see Bores-Rangel et al., 1990), and career decision profile (Jones, 1989; Jones and Lohmann, 1998). Personality development was measured with several instruments including the *Bem Sex Role Inventory* (Bem, 1974; Lenney, 1991), a revised version of *extended objective measure of ego identity status* (Grotevant and Adams, 1984), Rosenberg's *stability of self* (Alsaker and Olweus, 1986), *the self-perception profile for adolescents* (Wichstrøm, 1995), *state–trait Anger expression inventory* (Spielberger et al., 1999), *Barratt impulsiveness scale* (Patton et al., 1995), and Marlowe-Crown social desirability scale (Crowne and Marlowe, 1960). Finally, the questionnaire also included a variety of instruments assessing conduct, anxiety, and mood disorders. These mental-health related instruments were *Olweus' scale of antisocial behavior* (Olweus, 1989), *substance use* (Pape and Rossow, 2004), the *Bulimic investigatory test* (Henderson and Freeman, 1987), *the eating attitude test* (Garner and Garfinkel, 1979), the *Cantril ladder scale* (Cantril, 1965), *the UCLA loneliness scale* (Russell et al., 1978, 1980), *Hopkins symptom checklist* (Derogatis et al., 1974), and *depressive mood inventory* (Kandel and Davies, 1982).

Participants also were asked to respond to the item "*Have you ever tried to take your own life*," assessing attempted suicide, which could be answered in with a binary "yes" or "no." This item was used as the outcome variable for the classification task. The same item was examined at T2, allowing us to identify participants who had not reported a suicide attempt at T1 and are reporting a suicide attempt at T2.

Analysis

Model training and model selection

Utilizing the Gradient Boosting Algorithm (GBM; Friedman, 2001), we trained a binary classification model to identify suicide attempts at T1. The dataset was randomly divided into training (70%) and testing (30%) subsets. We fine-tuned the GBM algorithm with random search on the training dataset and employed a 10-fold Cross-Validation (CV) method to assess the performance of the models. The search algorithm was optimized to select models with highest Area Under Precision-Recall Curve (AUPRC). This metric is considered less biased than Area Under the Curve (AUC) or misclassification rate, especially when outcomes are rare and severely imbalanced (Davis and Goadrich, 2006; Chicco, 2017). We chose the model showcasing the highest AUPRC and evaluated its performance on the testing subset. In addition to AUPRC, we also analyzed the Receiver Operating Characteristic (ROC) curve and reported the AUC of the chosen model to make our results comparable with the literature. Note that the procedure of model training and model selection is exclusively conducted on T1. As detailed below, T2 is only used as a follow-up to examine the risk of FP and TN classification groups.

Classification based on different specificity thresholds

The machine learning classification model assigns a risk score for each subject in the test dataset, which can range from 0 to 1. The higher the estimated suicide risk score, the higher the chance of a past suicide attempt. Classification can be performed based on any chosen threshold value in this range, resulting in different rates of FP, TN, and True Positives (TP). The higher the cutoff value for classification, the less likely it is that individuals are falsely classified as positive. This, however, comes at the cost of misclassifying a higher proportion of true positive individuals who were assigned a lower risk score, thus increasing the False Negative (FN) group. We used the adjROC R package (Haghish, 2022a) to perform the classification for cutoff values corresponding to specificity levels ranging from 0.60 to 0.975 and accordingly, for each level we identified TP, TN, and FP. Crucially, all classifications were made based on the T1 data of the testing dataset only. In other words, using the selected model and T1 test dataset, we classified the test sample for a range of cutoff values, gradually increasing the specificity of the classification model and subsequently, identified the individuals which would be classified as TP, FP, and FN at each specificity threshold. Specifically, the T2 dataset was only used to examine the prevalence of suicide attempts among these classification groups. The procedure of model training, model selection, classification for a range of specificity values, and estimating relative risk for the first-time attempters at T2 is shown in Figure 2.

Next, we assessed the ratios with which FP and TN reported a suicide attempt for the first time at T2. We calculated the Relative Risk ($RR_{FP/TN}$) as shown below. Fisher's exact test of count data was used to evaluate whether the relative risk values were significantly greater than 1.0. Finally, we calculated the $RR_{FP/TN}$ as a function of the level of specificity to examine whether the relative risk increases for higher values of estimated risk, corresponding to higher specificity cutoff for the model as well.

$$RR_{FP/TN} = \frac{suicide\ prevalence\ at\ T2\ among\ FP}{suicide\ prevalence\ at\ T2\ among\ TN}$$

Missing data imputation

Missing observations in the predictors at the first time point were imputed using the mlim R package (Haghish, 2022b). The mlim package applies machine learning algorithms for missing data imputation and can handle mixed data types with complex interactions. This imputation algorithm is shown to result in lower imputation error compared to standard statistical procedures (Haghish, 2023). The outcome variable (i.e., suicide attempt) was separated from the dataset prior to the imputation and thus, missing data on the outcome variables were removed. After the imputation, the outcome variable was reattached to the dataset.

Results

In the first time point, 7.52% of the items were missing and therefore were imputed prior to the analysis. A previous suicide attempt at T1 was reported by 8.43% ($n = 913$) of the adolescents, of which 37.79% ($n = 345$) were boys and 62.21% ($n = 568$) were girls. Of the reported suicide attempts, 57.61% were by adolescents in senior high school (above 15 years old) and the rest (42.39%) in junior high school. Fine-tuning the algorithms, the best GBM model reached

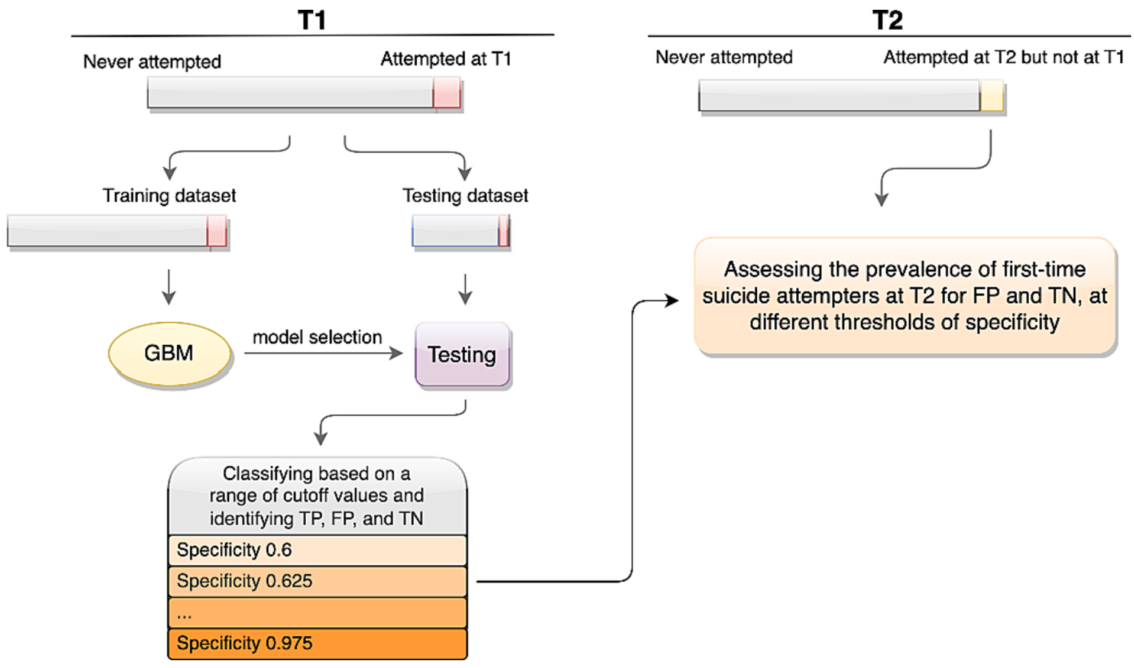


FIGURE 2
The study procedure.

AUPRC of 50.51% and AUC of 88.58%. Analysis of the model suggested a cutoff of 0.0587, resulting in sensitivity of 0.854 and specificity of 0.758. This cutoff value is shown in Figure 3 with a dotted line, alongside the histogram of the estimated suicide risk scores for the test dataset. Further analysis using the adjROC R package estimated that cutoff values ranging from 0.0346 to 0.286 would correspond to specificity values ranging from 0.60 to 0.975, respectively.

False positives' relative risk

At T2, 156 individuals reported their first suicide attempt. We determined which of these individuals were in the TN or FP groups based on different specificity thresholds. As depicted in Figure 4, we computed the $RR_{FP/TN}$ corresponding to rising specificity levels. This figure also displays the regression line and its 95% confidence interval for various specificity values. As shown in Figure 4, the relative risk spans from 2.96 for a specificity of 0.60 to 7.22 at a specificity of 0.975. Fisher's test indicated that both risks are significantly higher than 1.0 ($RR_{specificity=0.6} = 2.96$, 95% $CI = 1.74$ – Inf, $p = 0.0002$ and $RR_{specificity=0.975} = 7.22$, 95% $CI = 3.06$ – Inf, $p = 0.0002$). Moreover, as shown in Figure 4, the specificity threshold related to the FP's relative risk significantly increased with increasing specificity [*Adjusted* $R^2 = 0.865$, $F\ (1, 14) = 96.750$, $p < 0.0001$].

Similarity of false positives and true positives

In Figure 5, we plotted the normalized average severity of symptoms of depression, suicidal ideation, general anxiety, perceived loneliness, perceived personal problems (evaluated with an item asking "*do you have a personal problem that you need help with*"), and frequency of smoking among FP and TP for specificity levels ranging from 0.60 to 0.975. Apart from smoking, the average scores for the other variables in the FP and TP groups were more similar for higher levels of specificity. This was especially pronounced when specificity was above 0.9. In Figure 5, for both TP and FP, smoking is frequent. However, for high-risk adolescents, where specificity was above 0.9, smoking behavior appeared to be even more frequent.

To examine the third hypothesis, we modeled the difference between the two curves for each scale, which generally resulted in a linear trend, diminishing with increasing specificity. Apart from the smoking frequency that had a negligible reverse trend from specificity of 0.9–0.975 [a*djusted* $R^2 = -0.016$, $F\ (1, 36) = 0.42$, $p = 0.52$], other symptoms conformed to the anticipated trend, such as depression [*adjusted* $R^2 = 0.969$, $F\ (1, 36) = 1163.0$, $p < 0.0001$], suicide ideation [*adjusted* $R^2 = 0.963$, $F\ (1, 36) = 974.2$, $p < 0.0001$], anxiety [a*djusted* $R^2 = 0.959$, $F\ (1, 36) = 870.6$, $p < 0.0001$], loneliness [a*djusted* $R^2 = 0.797$, $F\ (1, 36) = 146.0$, $p < 0.0001$], and personal problems [*adjusted* $R^2 = 0.954$, $F\ (1, 36) = 751.0$, $p < 0.0001$].

Discussion

Within the context of machine-learning-based suicide attempt classification, we investigated three hypotheses concerning the suicide attempt risk of the FP group. Assuming the classification is executed by a highly precise model, our first hypothesis posited that, in comparison to the TN group, the FP group would exhibit a substantially elevated risk of suicide attempts over a two-year span. The fine-tuned model reached AUC of 88.58%, meeting the precondition that the classification model is highly accurate

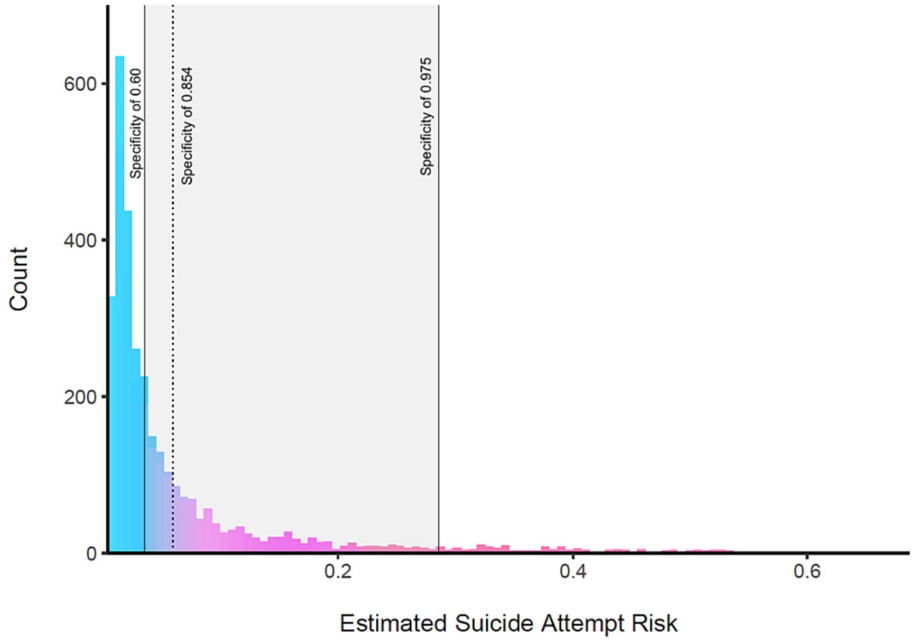


FIGURE 3
The estimated suicide attempt risk of the test dataset at T1. The gray rectangle indicates the range of estimated suicide attempt risk that corresponds to specificity levels ranging from 0.60 to 0.975, showing risk values ranging from 0.0346 to 0.286. As higher risk thresholds (values) for classification are selected, the specificity of the model also increases accordingly.

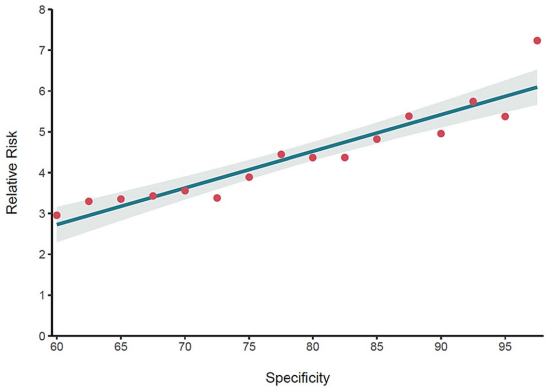


FIGURE 4
The relative risk of false positives at T2 for different specificity values. The red points represent the relative risk of the FP group at T2, indicating that FP are at much higher risk of attempting suicide compared to TN. The regression line indicates that the relative risk of the FP increases as a higher specificity is selected for the model.

(Šimundić, 2009). The results supported this hypothesis. In the two years after the initial assessment, the FP group's self-reported suicide attempts' prevalence was approximately 3 to 7 times that of the TN group. This statistically significant rise in relative risk indicates that a notable portion of the FP group may contemplate suicide in the future, particularly if the classification is made at a high specificity threshold. The second hypothesis suggested that the relative risk for the FP group would increase when classifications are set at higher specificity thresholds. The linear regression analysis showed a significant linear trend in support of this hypothesis. Finally, the third hypothesis proposed that as classification is made at higher specificity thresholds, the severity of the FP group's suicide attempt risk factors would more closely resemble those of the TP group. Apart from frequency of smoking, a clear trend was observed for other mental health indicators such as depression, anxiety, suicidal ideation, loneliness and personal problems, which supported this hypothesis. Overall, these findings are in line with Haghish and Czajkowski (2023) argument that the FP group can be conceptualized as a suicide risk group and a relevant target group for a suicide intervention program rather than as a mere classification error.

There are several reasons why the FP group exhibited an elevated risk of suicide attempt in our study. First, our machine learning model was based on a large number of psychological, sociological, and environmental risk factors that are likely to be persistent and might have a causal influence on the development of suicidal behavior (Lohner and Konrad, 2006; Greening et al., 2008; Darke et al., 2010; Toprak et al., 2011; Lewis et al., 2014; Carballo et al., 2020; Haghish, 2023). Depression, anxiety, substance use, suicidal ideation, and other mental health risk factors are likely to endure over time (Caspi and Moffitt, 2018), keeping the FP at a higher risk of suicide attempt in the future. We showed that this effect would increase as a function of specificity. Moreover, as the model becomes more accurate in identifying TP and TN, the risk for those identified as FP also increases; presumably, because they were expected to have more similar patterns or levels on the risk indicators as the TP. However, we did not examine the above question, which should be addressed by future studies and, ideally, with even larger datasets. This type of effect can also be observed in logistic regression models, which utilize fewer predictors and assume monotonic relationships between the predictors and the outcome variable. In contrast, machine learning models do not assume such relationships and search for patterns in

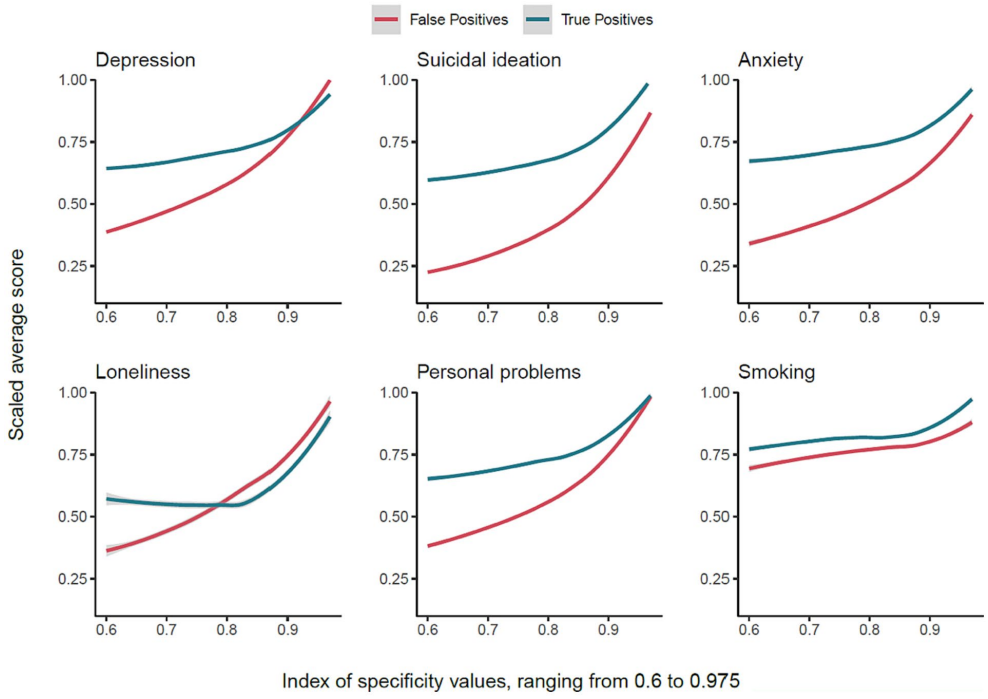


FIGURE 5
The mean score of FP increases alongside specificity and becomes similar to TP.

predictors, as well as interactions between them, in order to improve classification accuracy. Therefore, by increasing the specificity one can also assume that FP will have similar patterns to TP in their responses.

Methodological and clinical implications

We showed that false positives - in the context of a machine learning retrospective suicide attempts classification model - can have a different interpretation than that usually ascribed. This finding is noteworthy from a methodological as well as a clinical perspective. From a methodological perspective, our results suggest that we might need a fairer method to evaluate machine learning model performance whenever FP are expected to be at high risk of developing the outcome. In short, when such a model is used in mental health settings, rather than punishing the model for its FP error, it should be credited for identifying individuals at risk, as long as such individuals are clinically relevant. This would seem a fairer and more optimistic way to evaluate the model performance rather than pessimistically consider all FP as sheer errors. In addition to common model performance procedures that are centered on misclassified groups, if the identified FP are within the conditions listed in Figure 1, the model evaluation could also be done based on clinical relevance. Such an approach clearly requires research to define "clinical relevance" for different health problems as well as estimating the relative risk of FP in different contexts rather than solely underscoring a correct classification. In the context of a hypothetical intervention and when the model's accuracy is high and the classification is based on a high specificity, the individuals labeled as positive (whether TP or FP) are likely to benefit from the intervention. In this case, identifying individuals at high risk of becoming suicidal is clinically relevant for a prevention program.

In addition to severe symptoms of mental health problems, there are two other reasons why FP might be a relevant target group for a suicide prevention program. On the one hand, offering aid to individuals that already have attempted suicide does not guarantee a successful treatment, since there is little evidence in favor of effectiveness of suicide intervention programs for clinical samples (Large, 2018; Fox et al., 2020). Instead, preventing the development of suicidal behavior has been emphasized in recent studies as a better solution to reducing suicide prevalence in the population (Carter and Spittal, 2018; Haghish and Czajkowski, 2023). On the other hand, empirical evidence shows that even in Western countries such as Finland, Norway, and United States, most of adolescents' suicide attempts might go undetected. Thus, they might not receive the needed professional mental health support before or after attempting suicides (Suominen et al., 2004; Olfson et al., 2012; Haghish, 2023). In the United States, for instance, college counseling centers have reported that only 19% of the students who died of suicide have been in contact with the counseling centers that are instructed to provide suicide first-aid (Gallagher, 2009). Therefore, identifying adolescents who are at high risk of becoming suicidal in future might be an indispensable step toward suicide prevention. Toward this end, machine learning can provide reliable suicide risk estimations, which can help us identify risk groups that need attention. As shown in our results, such estimated risk scores are indicative of future suicide attempt risk, even when the machine learning model is trained with retrospective data.

Limitations and strengths

This study has several limitations that warrant attention. Primarily, suicide attempts were measured using a sole self-report item, leaving questions about the intensity and sincerity of these attempts. Nevertheless, this limitation does not undermine our conclusions which highlight that adolescents, even if inaccurately labeled as positive by a precise model, are at heightened risk of attempting suicide in the near future. Should the model incorporate more nuanced features reflecting the severity of suicide attempts, the relative risk associated with false positives is expected to escalate due to refined accuracy in risk estimation. In other words, the higher the model's accuracy, the more reliable its risk predictions, irrespective of its classification correctness. Moreover, a binary classification is useful for identifying *who should receive help* and not *when the individual may attempt suicide*. Nonetheless, as previously mentioned, this method can play a pivotal role in prevention. By recognizing adolescents who are on the verge of developing suicidal tendencies, timely interventions can be administered, potentially averting tragic outcomes. Our study has also several notable strengths. Firstly, it takes a critical perspective on the common practice of suicide attempt classification with machine learning, shedding light on its inherent limitations. Furthermore, our findings accentuate the clinical significance of the FP group under the aforementioned preconditions, which merits more attention from future research. Finally, this study leverages a large population-representative longitudinal data from Norwegian adolescents, which helped to train an accurate model.

Conclusion

We posited that the focus of suicide attempt classification should expand beyond those who have already attempted suicide to also encompass those poised to exhibit suicidal tendencies in the future. Notably, our findings suggest that it's plausible to pinpoint both groups, if a machine-learning model for classifying suicide attempts integrates a multitude of psycho-socio-environmental risk factors, and achieves commendable accuracy and specificity. In other words, the more reliable the estimated suicide attempt risk, the more the risk should be taken seriously, even for misclassified adolescents. Additionally, achieving this would necessitate estimations concerning the fraction of FP group likely to undertake a suicide attempt or exhibit suicidal tendencies, warranting further empirical research. Furthermore, we argued that supplementary performance metrics could be incorporated to consider the potential risk or clinical relevance of FP. Our results should be taken into account by future suicide intervention programs that intend to use survey data and machine learning classification algorithms to identify at-risk individuals. As this is the first study to examine the claim that FP can be conceptualized as a risk group, there is a clear need for investigating further whether these results are replicable and can be extended to machine learning classification or prediction models of other mental health outcomes.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: data from Young in Norway study was analysed. Researchers can apply for access to the data via https://ung-i-norge.no/. Requests to access these datasets should be directed to https://ung-i-norge.no/.

Ethics statement

The studies involving humans were approved by the Ethical committee at the department of psychology, university of Oslo. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

EFH developed the idea, carried out the analysis, and wrote the draft. BL and NC revised the manuscript. All authors contributed to the article and approved the submitted version.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alsaker, F., and Olweus, D. (1986). Assessment of global negative self-evaluations and perceived stability of self in Norwegian preadolescents and adolescents. *J. Early Adolesc.* 6, 269–278. doi: 10.1177/0272431686063005

Bem, S. L. (1974). The measurement of psychological androgyny. *J. Consult. Clin. Psychol.* 42, 155–162. doi: 10.1037/h0036215

Bores-Rangel, E., Church, A. T., Szendre, D., and Reeves, C. (1990). Self-efficacy in relation to occupational consideration and academic performance in high school equivalency students. *J. Couns. Psychol.* 37, 407–418. doi: 10.1037/0022-0167.37.4.407

Bourdieu, P. (2018). "Distinction a social critique of the judgement of taste" in *Inequality Classic Readings in Race, Class, and Gender* (New York: Routledge), 287–318.

Burke, T. A., Jacobucci, R., Ammerman, B. A., Alloy, L. B., and Diamond, G. (2020). Using machine learning to classify suicide attempt history among youth in medical care settings. *J. Affect. Disord.* 268, 206–214. doi: 10.1016/j.jad.2020.02.048

Cantril, H. (1965). *The pattern of human concern*. New Brunswick, NJ: Rutgers University Press.

Carballo, J., Llorente, C., Kehrmann, L., Flamarique, I., Zuddas, A., Purper-Ouakil, D., et al. (2020). Psychosocial risk factors for suicidality in children and adolescents. *Eur. Child Adolesc. Psychiatry* 29, 759–776. doi: 10.1007/s00787-018-01270-9

Carter, G., and Spittal, M. J. (2018). Suicide risk assessment. *Crisis* 39, 229–234. doi: 10.1027/0227-5910/a000558

Caspi, A., and Moffitt, T. E. (2018). All for one and one for all: mental disorders in one dimension. *Am. J. Psychiatr.* 175, 831–844. doi: 10.1176/appi.ajp.2018.17121383

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *Bio Data Min.* 10, 1–17. doi: 10.1186/s13040-017-0155-3

Crowne, D. P., and Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *J. Consult. Psychol.* 24, 349–354. doi: 10.1037/h0047358

Darke, S., Torok, M., Kaye, S., and Ross, J. (2010). Attempted suicide, self-harm, and violent victimization among regular illicit drug users. *Suicide Life Threat. Behav.* 40, 587–596. doi: 10.1521/suli.2010.40.6.587

Davis, J., and Goadrich, M. (2006). "The relationship between Precision-Recall and ROC curves" in *Proceedings of the 23rd International Conference on Machine Learning*. New York: Association for Computing Machinery. 233–240.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., and Covi, L. (1974). The Hopkins Symptom Checklist (HSCL): a self-report symptom inventory. *Behav. Sci.* 19, 1–15. doi: 10.1002/bs.3830190102

Fox, K. R., Huang, X., Guzmán, E. M., Funsch, K. M., Cha, C. B., Ribeiro, J. D., et al. (2020). Interventions for suicide and self-injury: a meta-analysis of randomized controlled trials across nearly 50 years of research. *Psychol. Bull.* 146, 1117–1145. doi: 10.1037/bul0000305

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., et al. (2017). Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol. Bull.* 143:187. doi: 10.1037/bul0000084

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232.

Gallagher, R. P. (2009). *National survey of counseling center directors 2009*. Alexandria, VA: International Association of Counseling Service. Available at: https://d-scholarship.pitt.edu/28170/1/survey_2009.pdf

Garner, D. M., and Garfinkel, P. E. (1979). The Eating Attitudes Test: an index of the symptoms of anorexia nervosa. *Psychol. Med.* 9, 273–279. doi: 10.1017/S0033291700030762

Greening, L., Stoppelbein, L., Fite, P., Dhossche, D., Erath, S., Brown, J., et al. (2008). Pathways to suicidal behaviors in childhood. *Suicide Life Threat. Behav.* 38, 35–45. doi: 10.1521/suli.2008.38.1.35

Grotevant, H. D., and Adams, G. R. (1984). Development of an objective measure to assess ego identity in adolescence: validation and replication. *J. Youth Adolesc.* 13, 419–438. doi: 10.1007/BF02088639

Haghish, E. F. (2022a). *AdjROC: computing sensitivity at a fix value of specificity and vice versa (0.2.0) [Computer software]*. Available at: https://CRAN.R-project.org/package=adjROC

Haghish, E. F. (2022b). *mlim: single and multiple imputation with automated machine learning (0.3) [Computer software]*. Available at: https://CRAN.R-project.org/package=mlim

Haghish, E. F. (2023). *mlim: single and multiple imputation with automated machine learning [GItHub Repository]*. Available at: https://github.com/haghish/mlim

Haghish, E. F., and Czajkowski, N. O. (2023). Reconsidering false positives in machine learning binary classification models of suicidal behavior. *Curr. Psychol.* (2023). doi: 10.1007/s12144-023-05174-z

Haghish, E. F., Czajkowski, N. O., and von Soest, T. (2023). Predicting suicide attempts among Norwegian adolescents without using suicide-related items: a machine learning approach Frontiers in Psychiatry [manuscript submitted for peer-review].

Healy, B. C. (2021). Machine and deep learning in MS research are just powerful statistics–No. *Mult. Scler. J.* 27, 663–664. doi: 10.1177/1352458520978648

Henderson, M., and Freeman, C. (1987). A self-rating scale for bulimia the BITE. *Br. J. Psychiatry* 150, 18–24. doi: 10.1192/bjp.150.1.18

Jones, L. K. (1989). Measuring a three-dimensional construct of career indecision among college students: a revision of the vocational decision scale: the career decision profile. *J. Couns. Psychol.* 36:477. doi: 10.1037/0022-0167.36.4.477

Jones, L. K., and Lohmann, R. C. (1998). The career decision profile: using a measure of career decision status in counseling. *J. Career Assess.* 6, 209–230. doi: 10.1177/106907279800600207

Kandel, D. B., and Davies, M. (1982). Epidemiology of depressive mood in adolescents: an empirical study. *Arch. Gen. Psychiatry* 39, 1205–1212. doi: 10.1001/archpsyc.1982.04290100065011

Large, M. M. (2018). The role of prediction in suicide prevention. *Dialogues Clin. Neurosci.* 20, 197–205. doi: 10.31887/DCNS.2018.20.3/mlarge

Lenney, E. (1991). "Sex roles: the measurement of masculinity, femininity, and androgyny" in *Measures of personality and social psychological attitudes*. eds. J. P. Robinson, P. R. Shaver and L. S. Wrightsman (San Diego: Academic Press), 573–660.

Lewis, A. J., Bertino, M. D., Bailey, C. M., Skewes, J., Lubman, D. I., and Toumbourou, J. W. (2014). Depression and suicidal behavior in adolescents: a multi-informant and multi-methods approach to diagnostic classification. *Front. Psychol.* 5:766. doi: 10.3389/fpsyg.2014.00766

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., and Tischer, T. (2022). Machine learning and conventional statistics: making sense of the differences. *Knee Surg. Sports Traumatol. Arthrosc.* 30, 753–757. doi: 10.1007/s00167-022-06896-6

Linthicum, K. P., Schafer, K. M., and Ribeiro, J. D. (2019). Machine learning in suicide science: applications and ethics. *Behav. Sci. Law* 37, 214–222. doi: 10.1002/bsl.2392

Lohner, J., and Konrad, N. (2006). Deliberate self-harm and suicide attempt in custody: distinguishing features in male inmates' self-injurious behavior. *Int. J. Law Psychiatry* 29, 370–385. doi: 10.1016/j.ijlp.2006.03.004

Marjoribanks, K. (1987). Ability and attitude correlates of academic achievement: family-group differences. *J. Educ. Psychol.* 79, 171–178. doi: 10.1037/0022-0663.79.2.171

Olfson, M., Marcus, S. C., and Bridge, J. A. (2012). Emergency treatment of deliberate self-harm. *Arch. Gen. Psychiatry* 69, 80–88. doi: 10.1001/archgenpsychiatry.2011.108

Olweus, D. (1989). "Prevalence and incidence in the study of antisocial behavior: definitions and measurements" in *Cross-national research in self-reported crime and delinquency*. ed. M. W. Klein (Dordrecht: Springer), 187–201.

Pape, H., and Rossow, I. (2004). "Ordinary" people with "normal" lives? A longitudinal study of ecstasy and other drug use among Norwegian youth. *J. Drug Issues* 34, 389–418. doi: 10.1177/002204260403400207

Patton, J. H., Stanford, M. S., and Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *J. Clin. Psychol.* 51, 768–774. doi: 10.1002/1097-4679(199511)51:6<768::AID-JCLP2270510607>3.0.CO;2-1

Russell, D., Peplau, L. A., and Cutrona, C. E. (1980). The revised UCLA Loneliness Scale: concurrent and discriminant validity evidence. *J. Pers. Soc. Psychol.* 39:472. doi: 10.1037/0022-3514.39.3.472

Russell, D., Peplau, L. A., and Ferguson, M. L. (1978). Developing a measure of loneliness. *J. Pers. Assess.* 42, 290–294. doi: 10.1207/s15327752jpa4203_11

Šimundić, A.-M. (2009). Measures of diagnostic accuracy: basic definitions. *Ejifcc* 19:203.

Spielberger, C. D., Sydeman, S. J., Owen, A. E., and Marsh, B. J. (1999). *Measuring anxiety and anger with the State-Trait Anxiety Inventory (STAI) and the State-Trait Anger Expression Inventory (STAXI)*. Mahwah: Lawrence Erlbaum Associates Publishers.

Storvoll, E. E., and Wichstrøm, L. (2002). Do the risk factors associated with conduct problems in adolescents vary according to gender? *J. Adolesc.* 25, 183–202. doi: 10.1006/jado.2002.0460

Suominen, K., Isometsä, E., Martunnen, M., Ostamo, A., and Lönnqvist, J. (2004). Health care contacts before and after attempted suicide among adolescent and young adult versus older suicide attempters. *Psychol. Med.* 34, 313–321. doi: 10.1017/S0033291703008882

Toprak, S., Cetin, I., Guven, T., Can, G., and Demircan, C. (2011). Self-harm, suicidal ideation and suicide attempts among college students. *Psychiatry Res.* 187, 140–144. doi: 10.1016/j.psychres.2010.09.009

van Vuuren, C., van Mens, K., de Beurs, D., Lokkerbol, J., van der Wal, M., Cuijpers, P., et al. (2021). Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey. *J. Affect. Disord.* 295, 1415–1420. doi: 10.1016/j.jad.2021.09.018

Walsh, C. G., Ribeiro, J. D., and Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clin. Psychol. Sci.* 5, 457–469. doi: 10.1177/2167702617691560

Wichstrøm, L. (1995). Harter's Self-Perception Profile for Adolescents: Reliability, validity, and evaluation of the question format. *J. Pers. Assess.* 65, 100–116. doi: 10.1207/s15327752jpa6501_8

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 12 October 2023
DOI 10.3389/fpsyg.2023.1273582

OPEN ACCESS

EDITED BY
Sai-fu Fung,
City University of Hong Kong,
Hong Kong SAR, China

REVIEWED BY
Dimitrios Stamovlasis,
Aristotle University of Thessaloniki, Greece
Kosuke Kawai,
University of California,
Los Angeles, United States

*CORRESPONDENCE
Georgios Sideridis
georgios.sideridis@gmail.com

RECEIVED 06 August 2023
ACCEPTED 22 September 2023
PUBLISHED 12 October 2023

CITATION
Sideridis G and Jaffari F (2023) Identifying person misfit using the person backward stepwise reliability curve (PBRC).
Front. Psychol. 14:1273582.
doi: 10.3389/fpsyg.2023.1273582

Identifying person misfit using the person backward stepwise reliability curve (PBRC)

Georgios Sideridis[1,2]* and Fathima Jaffari[3]

[1]Boston Children's Hospital, Harvard Medical School, Boston, MA, United States, [2]National and Kapodistrian University of Athens, Athens, Greece, [3]Education and Training Evaluation Commission, Riyadh, Saudi Arabia

The goal of the present study was to propose a visualization of aberrant response patterns based on the idea put forth by the Cronbach-Mesbach curve. First, an index of person reliability is developed using the K-R 20 formula followed by a backward stepwise procedure in which one person at a time is deleted from the model. Observations for which reliability is no longer monotonically increasing suggest that they are candidates for aberrant responding. Using data from the quantitative domain of a national aptitude test the proposed visualization technique was demonstrated. The external validity of the procedure was tested by contrasting the person fit reliability estimates with those derived from other indices of aberrant responding such as the Ht. Results indicated that individuals not covarying with other individuals concerning their response patterns and concordance to the measurement of a unified latent trait were identified by both the present procedure and Ht and U3 at a rate of 100%. By plotting those individuals using Person Response Curves (PRCs) results confirmed the lack of monotonicity in the relationship between item difficulty and person skill. Consequently, results confirm the usefulness of the present methodology as an index for identifying responders who manifest themselves with aberrant responses and who are not conducive to the measurement of the latent trait.

KEYWORDS

person reliability, K-R 20, aberrant responding, person fit, visual analysis

1. Introduction

When individuals take a test, several processes are operative that may affect the way of responding which may result in the provision of invalid results. This notion of behaving in aberrant and unexpected ways represents a serious threat to the validity of test results with significant implications for both the person and the instrument (Little and Moore, 2013; Ferro and Beaton, 2016) as test scores include construct-irrelevant variance (Messick, 1995). At the personal level, individuals may obtain results substantially higher (as in cheating-see Cizek, 1999) or lower (as in being inattentive and careless, Meade and Craig, 2012) with significant implications for placement, selection, academic and job opportunities, etc.

Types of aberrant response patterns may involve random guessing (Lord, 1964), withdrawal (Ward et al., 2017), carelessness (Rios et al., 2017), speeding (Wise and Kong, 2005), rapid guessing (Deribo et al., 2021), inattentiveness (McKay et al., 2018), the presence of acquiescence (Plieninger and Heck, 2018), faking (Paulhus, 1991), social desirability (Leite and Cooper, 2010), recall biases (Barry, 1996), random responding (Cook et al., 2016), non-responding (Groves, 2006), ineffective strategy use (e.g., skipping items), the engagement of response sets (Müller

et al., 2015), extreme responding (Meisenberg and Williams, 2008), response drifting (Drasgow and Parsons, 1983), insufficient effort (Hong et al., 2019), insufficient responding (Bowling et al., 2016), etc. Regardless of whether such behaviors are intentional or not, they have a major impact on the reliability and validity of the obtained scores. Thus, it is important to have tools to identify aberrant responses so that processes may be put in place to address the validity of test scores as they reflect the person or the instrument in total and likely represent a major threat to validity (van Laar and Braeken, 2022).

1.1. Reliability in measurement and aberrant responding

Ultimately, the quality of measurement is expressed by the ability of an instrument to provide measurements that are accurate, precise, and repeatable. This concept of reliability of measurement is most often discussed and estimated using information derived from a sample on a scale's components, such as the items. One of the proponents of internal consistency reliability was Cronbach (1951) who also proposed the alpha coefficient as a reflection of the strength of the relationships between a set of items and the measured construct, assuming unidimensionality. Alpha is expressed using the following formula:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i}{\sigma^2_{y_i}}\right) \quad (1)$$

With K being the number of the items in the scale; and σ_i *and* $\sigma^2_{y_i}$ the item's variances and total variance, respectively. As a means to improve the internal consistency of a measure that does not reach acceptable standards, an item analysis methodology termed "reliability if item deleted" has been proposed so that one item at a time is excluded and alpha is re-expressed with the remaining items. The value of alpha is then evaluated with and without the removed item and decisions regarding internal consistency and unidimensionality are based on those estimates.

Mesbah (2010) put forth a graphical method using the logic of "alpha if item deleted" for evaluating the unidimensionality of a set of items. This stepwise method engages the "Backward Reliability Curve – *BRC*" with alpha being graphed after each successive step. Initially, the value of alpha is calculated using all items of a latent variable. After that, one item would be removed at a time with the value of alpha being re-estimated with the remaining items. The selection of the item in a stepwise fashion is based on the one that maximizes alpha if the item is deleted. Thus, the stepwise method concludes when only two items remain. Based on Classical Test Theory (CTT) and the Spearman–Brown formula, adding more items to the scale increases its reliability, thus a monotonically increasing BRC is expected when all items contribute to the formation of a unidimensional latent variable.

The present study extends the idea of the BRC at the person level by graphing a scale's reliability using a person-deleted stepwise procedure and plotting the reliability of a measure by examining how each person contributes to the measurement of a reliable unidimensional structure. In other words, the goal of the present graphical person-deleted alpha is to identify, and subsequently discard, individuals who behave in ways that the reliability of a measure is compromised. This procedure provides information about the sensitivity of the measure to individual responses by identifying individuals with aberrant response patterns that deviate markedly from the model's expectations (see Meijer, 1994). Thus, the original graphical method can be applied at the person level with the difference being that instead of removing \ adding one item at a time, we remove \ add one person at a time. Any decrease in the value of the reliability of the measure and the monotonic relationship expected by the BRC would be indicative of a person that is not constructive for measurement purposes or otherwise, that his/her response pattern reflects aberrant responding such as inattention or carelessness (Kam and Chan, 2018). To validate the proposed methodology, we employed a person-fit analysis with a known index that evaluates aberrant responding patterns. A substantial overlap in the selection of individuals who behave in unexpected ways following the Guttman pattern using the person BRC, and person fit statistics would provide evidence for the validity of the proposed methodology. Furthermore, by employing Person Response Curves (PRCs) the presence of aberrant responding will be evident in individuals whose curve does not conform to the descending trend as item difficulty increases. Thus, the goal of the present study was to introduce the Person Backward Reliability Curve (PBRC) and examine its criterion-related validity of selected misbehaving individuals in relation to the Ht index (Meijer and Sijtsma, 2001) and using Person Response Curves (PRCs).

2. Method

2.1. Participants and measure

Participants were $n = 82$ students who were part of a pilot study to evaluate general aptitude using the General Ability Test (GAT) which is a national criterion for university admission in Saudi Arabia. The quantitative domain utilized here was comprised of 44 items using a dichotomous scaling system. The quantitative domain assesses arithmetic, number sequence, analysis, logic, inductive reasoning, spatial ability relations, and visualization and is reflective of a single general dimension. In the present study we tested for the unidimensiionality of the measure by choosing among competing models using modern psychometrics.

2.2. Data analyzes

Three types of person-based analyzes for investigating aberrant response patterns were engaged, (a) the person backward reliability curve (PBRC), (b) the visual analysis of Person Response Curves (PRCs), and (c) the analysis of response vectors using person fit indices such as the Ht (Meijer and Sijtsma, 2001) and U3 (Van der Flier, 1982). The level of significance was set to 5% for a two-tailed test. In the presence of a family of tests (e.g., Table 1), we corrected for family-wise error using the Benjamini Hochberge corrective procedure. We opted against the popular Bonferroni procedure due to its conservatism and the fact that it does not adequately control for the false discovery rate (Holm, 1979; Nakagawa, 2004).

TABLE 1 Item fit statistics for quantitative domain, discrimination, and item difficulties.

Item No.	X^2	*d.f.*	Value of *p*	p-BH	a	*s.e.*	*b*	*s.e.*
6	6.520	1	0.011	0.294	2.440	1.050	−1.810	0.400
10	25.540	13	0.020	0.294	0.790	0.290	0.700	0.400
19	13.560	6	0.035	0.294	1.120	0.550	−2.230	0.840
8	6.230	2	0.044	0.294	2.120	0.890	−1.710	0.400
39	21.730	13	0.060	0.294	1.400	0.420	−0.330	0.210
20	17.610	10	0.062	0.294	1.570	0.450	−0.320	0.200
14	11.870	6	0.065	0.294	3.010	0.940	−0.670	0.160
5	23.610	15	0.072	0.294	0.560	0.260	−0.100	0.420
4	15.540	9	0.077	0.294	1.400	0.390	0.500	0.260
32	16.160	10	0.095	0.294	0.960	0.320	0.600	0.340
37	17.270	11	0.100	0.294	1.160	0.360	0.020	0.250
23	15.880	10	0.103	0.294	1.580	0.440	0.220	0.220
13	14.450	9	0.107	0.294	2.070	0.610	−0.570	0.180
24	21.970	15	0.108	0.294	0.590	0.270	−0.190	0.410
21	14.700	10	0.143	0.362	1.190	0.350	0.450	0.280
25	10.420	7	0.165	0.374	1.360	0.420	1.450	0.410
17	17.530	13	0.176	0.374	0.580	0.300	−1.420	0.730
2	16.090	12	0.187	0.374	1.160	0.350	0.400	0.280
27	14.890	11	0.187	0.374	0.990	0.380	−1.220	0.410
36	16.230	13	0.236	0.449	0.300	0.270	−3.190	2.870
16	12.290	10	0.266	0.461	0.970	0.380	−1.310	0.460
38	8.790	7	0.270	0.461	1.770	0.590	−1.140	0.260
12	16.570	14	0.279	0.461	0.580	0.270	1.110	0.650
31	12.530	11	0.327	0.510	1.900	0.520	−0.060	0.190
30	5.720	5	0.336	0.510	3.230	1.040	−0.660	0.150
29	12.200	11	0.350	0.512	1.560	0.460	−0.450	0.200
15	12.240	12	0.428	0.599	0.810	0.340	−1.240	0.500
3	11.020	11	0.443	0.599	1.140	0.340	0.820	0.330
1	12.900	13	0.457	0.599	1.010	0.370	−1.070	0.380
33	15.470	16	0.492	0.611	0.620	0.260	0.590	0.470
34	11.160	12	0.516	0.611	0.900	0.310	0.620	0.360
28	13.920	15	0.533	0.611	0.660	0.290	−0.750	0.440
7	14.680	16	0.550	0.611	0.600	0.270	−0.540	0.440
26	14.510	16	0.562	0.611	0.400	0.270	−1.700	1.180
22	8.690	10	0.563	0.611	1.860	0.520	−0.220	0.190
11	–	–	–	–	7.190	6.110	−1.440	0.200
35	7.730	10	0.656	0.692	1.320	0.420	−0.640	0.240
18	8.780	14	0.846	0.868	0.920	0.320	−0.400	0.290
9	3.800	9	0.924	0.924	2.250	0.620	0.120	0.190

p-BH are *p*-values corrected using the Benjamini-Hochberg correction; a, discrimination parameter; b, item difficulty; c.s.e.m, conditional standard error of measurement.

2.2.1. Backward reliability curve (BRC) and the person variant (PBRC)

The analysis based on the backward reliability curve originates from the work of Mesbah (2010) who attempted to graphically describe unidimensionality. He furthermore stated that a combination of items reflects a unidimensional construct if each item is related to the underlying latent dimension exclusively (Hamon and Mesbah, 2002). Furthermore, using Cronbach's alpha he suggested that the internal consistency reliability of a measure tends to increase with an increase in the number of items. Graphically speaking he proposed the Backward Reliability

Curve (BRC) that is being estimated in multiple steps with the first step including all items. Then at each subsequent step, one variable is removed from the model so that the variable selected is the one that results in the maximum value of Cronbach's alpha. Given that a monotonic relationship must exist between the number of items and alpha if an item is associated with a decrease in the curve, then that item is suspected that it does not contribute to the latent construct under evaluation. Under those lenses, items that are not associated with increases in the BRC, are candidates for exclusion.

In the present study, we propose two modifications to the BRC. First, by transposing items and columns, the BRC would be reflective of individuals who are constructive for measurement purposes, hence the term Personal Backward Reliability Curve (PBRC). Thus, individuals that lead to BRC decays are suspect and subject to removal. Second, we substituted Cronbach's alpha with the Kuder–Richardson estimation, which is appropriate for binary data (see Supplementary material on modification of CMC package functions). Consequently, the PBRC can utilize individuals who are only reflecting an increasing curve, thus, representing a more reliable measurement.

2.2.2. Ht and U3 person fit indices

The Ht coefficient, as presented by Meijer and Sijtsma (2001), is a measure used to quantify the extent to which data adhere to the Guttman model (Guttman, 1944; Meyer et al., 2013) for a single respondent in comparison to the other respondents within a given sample. The Ht coefficient is calculated by summing the covariances between the respondent's responses and the responses of the other respondents in the sample in the form of a covariance ratio as shown below:

$$H^T = \frac{\text{cov}\left(\boldsymbol{x}_n, \boldsymbol{r}_{(n)}\right)}{\text{cov}_{\max}\left(\boldsymbol{x}_n, \boldsymbol{r}_{(n)}\right)}, \quad (2)$$

With xn being the response vector for person n, and $r(n)$ being the response vector of total scores calculated from every participant in the sample except the xn person. Karabatsos (2003) suggested a cutoff value of <0.22 for Ht.

The maximum possible value of the Ht coefficient is 1, which indicates that the respondent's responses perfectly conform to the Guttman scale. A lower value of the Ht coefficient indicates that the respondent's responses are less consistent with the Guttman scale with values greater than 0.3 being suggestive of acceptable levels (Wongpakaran et al., 2019) or greater than 0.22 (Karabatsos, 2003). Simulation studies have shown that it has a high level of accuracy in detecting aberrant responses when applied to data with dichotomous response scales across different settings (Karabatsos, 2003; Dimitrov and Smith, 2006; Tendeiro and Meijer, 2014). Ht does not have a known theoretical distribution thus tests of inferential statistics cannot be conducted compared to other indices (e.g., lz*, Snijders, 2001; Magis et al., 2012) but given its efficacy in past research, it will be used as one of our two golden standards to determine the criterion validity of the proposed PBRC methodology.

The second person-fit index utilized, the U3 statistic, was developed by Van der Flier (1982) and was found to be the most accurate for the detection of random responding (Karabatsos, 2003) compared to all other tested indices ($n = 36$). Several studies confirmed the efficacy of U3 as an index of inattentive responding (e.g., Beck et al., 2019). The index reflects the ratio of the actual number of Guttman errors in a response pattern relative to the maximum number of errors using the log scale (Emons et al., 2005). It is being estimated as follows:

$$U3 = \frac{f\left(\mathbf{x}_n^*\right) - f\left(\mathbf{x}_n\right)}{f\left(\mathbf{x}_n^*\right) - f\left(\mathbf{x}_n^{'}\right)} \quad (3)$$

With $\mathbf{x}_n^*$ being the Guttman vector with correct responses for the easiest items in s_n, $\mathbf{x}_n^{'}$ the reversed Guttman vector with correct responses for the s_n hardest items, and $f\left(\mathbf{x}_n\right)$ being the summation $\sum_{i=1}^{I} x_{ni} \log\left[p_i / \left(1 - p_i\right)\right]$. In the Mousavi et al. (2019) study, the U3 index outperformed the Ht index across most conditions. Karabatsos suggested a cutoff value of 0.25 for U3 but Mousavi et al. (2019) challenged this cutoff value that was based on the standard normal and instead favored the value of p method and/or bootstrapping. All person fit indices were analyzed using the Perfit package (Tendeiro et al., 2016) in the R environment (R Core Team, 2017).

2.2.3. Analysis of person response curves (PRCs)

As an ancillary way of evaluating and validating a person's misfit, we will plot a person's proclivity to success using Person Response Curves (PRC). PRCs represent graphical means to evaluate the probability of a person's success on items of increasing difficulty. Thus, for any given individual, the expectation is that the curve will show a descending relationship with item difficulty by the use of an S-shaped curve. The curve is expected to start high as a person is likely successful on the easy items and is expected to gradually descend as the likelihood of correct responding goes down. Irregular PRCs would suggest that individuals are less successful on items that are within their level of ability and more successful on items that are out of reach, representing unexpected patterns more likely linked to inattention and/or cheating.

3. Results

3.1. Item response model for quantitative scale

A 2PL Item Response model was fit to the data and model fit was evaluated using descriptive fit indices and the RMSEA as well as the omnibus chi-square test. Results indicated acceptable model fit as the chi-square test was non-significant [χ^2 (702) = 751.598, $p = 0.95$]. Furthermore, the CFI and TLI were 0.936 and 0.932, respectively. Last, the RMSEA point estimate was 0.029 ($\text{RMSEA}_{95\%\text{CI}} = 0.000–0.046$). When contrasting the 2PL model to the fixed discrimination parameters model (Rasch), results indicated the superior fit of the 2PL model. Specifically, the Bayesian Information Criterion (BIC) values were 3732.72 for the 2PL model and 4073.19 for the Rasch model, suggesting the superiority of the former. Thus, collectively all information pointed to a good model fit using the 2PL model supporting the unidimensionality of the latent quantitative skills construct. Table 1 displays item-based parameters and item fit for the

instrument under study. Related to item misfit, all the corrected item-fit statistics based on the chi-square test suggested that items fit the premises of the Item Response Theory (IRT) model well and specifically the Guttman related pattern. Supplementary Figure S1 shows the Test Information Function (TIF) of the measure which peaked close to zero or slightly less than that and decays as it moves away further from mean theta, as expected with estimates deviating markedly from the mean and becoming less precise.

3.2. Person-based analyzes

3.2.1. Person backward reliability curve (PBRC) and person response curves (PRCs)

Figure 1 displays the proposed person backward reliability curve using fewer observations for illustration purposes. As shown in the figure, as participants are added to the measure so does internal consistency reliability which peaks at around 0.953 using the K-R formula. However, following that peak, the curve decays suggesting that the inclusion of specific individuals results in decrements in the model's estimated reliability. These observations were persons with ids 28, 5, 78, 23, 15, 20, 17, 9, 77, and 67. Thus, by merely using graphical means, these participants contribute amounts of error that are linked to decay in the measurement of internal consistency reliability. In other words, these participants are not contributing valuable information to the measure's reliability. Further analyzes of their response vectors highlight the possible causes for that misfit as highlighted by the PBRC.

Figure 2 displays the Person Response Curves (PRCs) for the 10 responders who were associated with decrements in the PBRC in Figure 1. As shown in the figure no participant displayed a PRC that was S-shaped with decays associated with decreases in item difficulty levels. As an example, the PRC of the first individual, id 28, displays a wave-like pattern with actual increases in item difficulty being associated with increases in the probability of success, which, as a pattern of behavior is against any of the premises of item response models. Person 28 had a theta estimate of 0.81 (S.E. = 0.308), thus, representing an above-average ability individual, who, however, was more successful on items beyond her/his ability level likely reflecting cheating; furthermore, this participant was unsuccessful on items within her/his ability level, likely reflecting inattention.

3.2.2. Person analysis of response vectors using Ht and U3

As mentioned above, for the analysis of response vectors, the Ht coefficient was utilized given its efficacy in past research (Karabatsos, 2003) to identify aberrant responders specifically linked to lucky guessing and cheating. Misfitted participants were flagged using cutoff values of 0.10 based on bootstrapping to simulate the sampling distribution of the Ht index with the current sample at the predetermined level of significance of 5% (Tendeiro et al., 2016; Mousavi et al., 2019). Figure 3, upper panel, displays the bootstrap distribution of Ht and its cutoff level of 0.10 (upper panel). Interestingly, below the cutoff Ht estimate of 0.10, there were 10 participants, which were exactly those identified using the PBRC. The only difference was in the ordering of participants Ht flagging in order of aberrance participants 78, 28, 5, 67, 23, 15, 20, 17, 77, and last, participant 9.

Similar results were observed with the use of U3. Using a value of p of 5%, the U3 index flagged 8 participants utilizing a cutoff value of 0.376 based on the bootstrap distribution (see Figure 3, lower panel). These participants and in the order of aberrance were ids: 5, 9, 15, 17, 20, 23, 28, and 78. Thus, all 8 flagged participants using U3 were also identified by the Cronbach-Mesbach curve, again supporting the criterion validity of the proposed PBRC at a level of 80% as two participants were not flagged using the alpha level of 5%.

4. Discussion and concluding remarks

The goal of the present study was to propose a visualization of aberrant response patterns based on the idea put forth by the

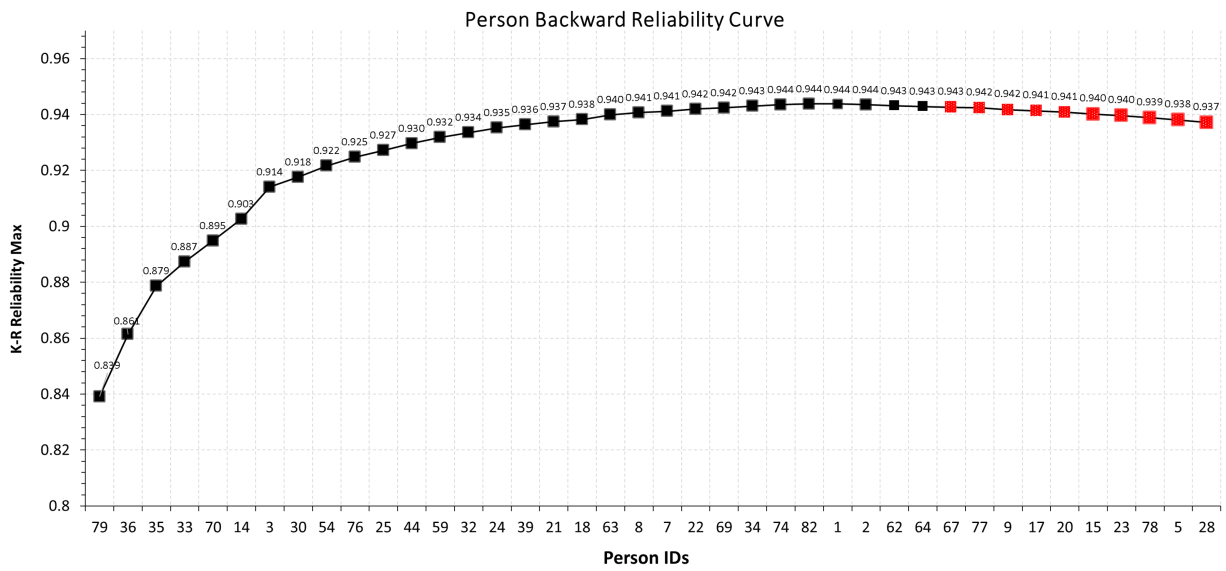


FIGURE 1
Modified Cronbach-Mesbach curve for the assessment of person reliabilities in relation to total person reliability using the K-R formula.

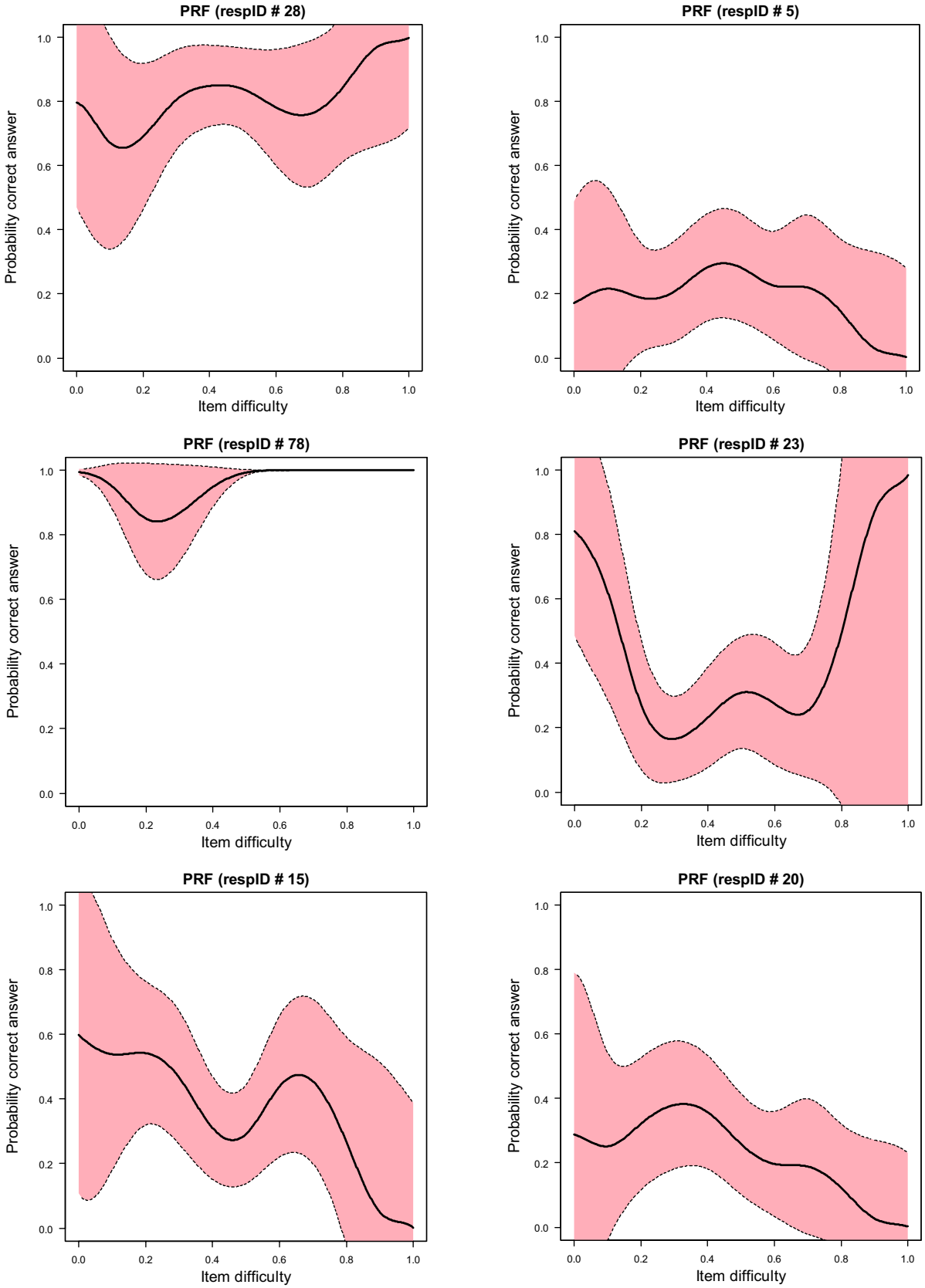


FIGURE 2 (Continued)

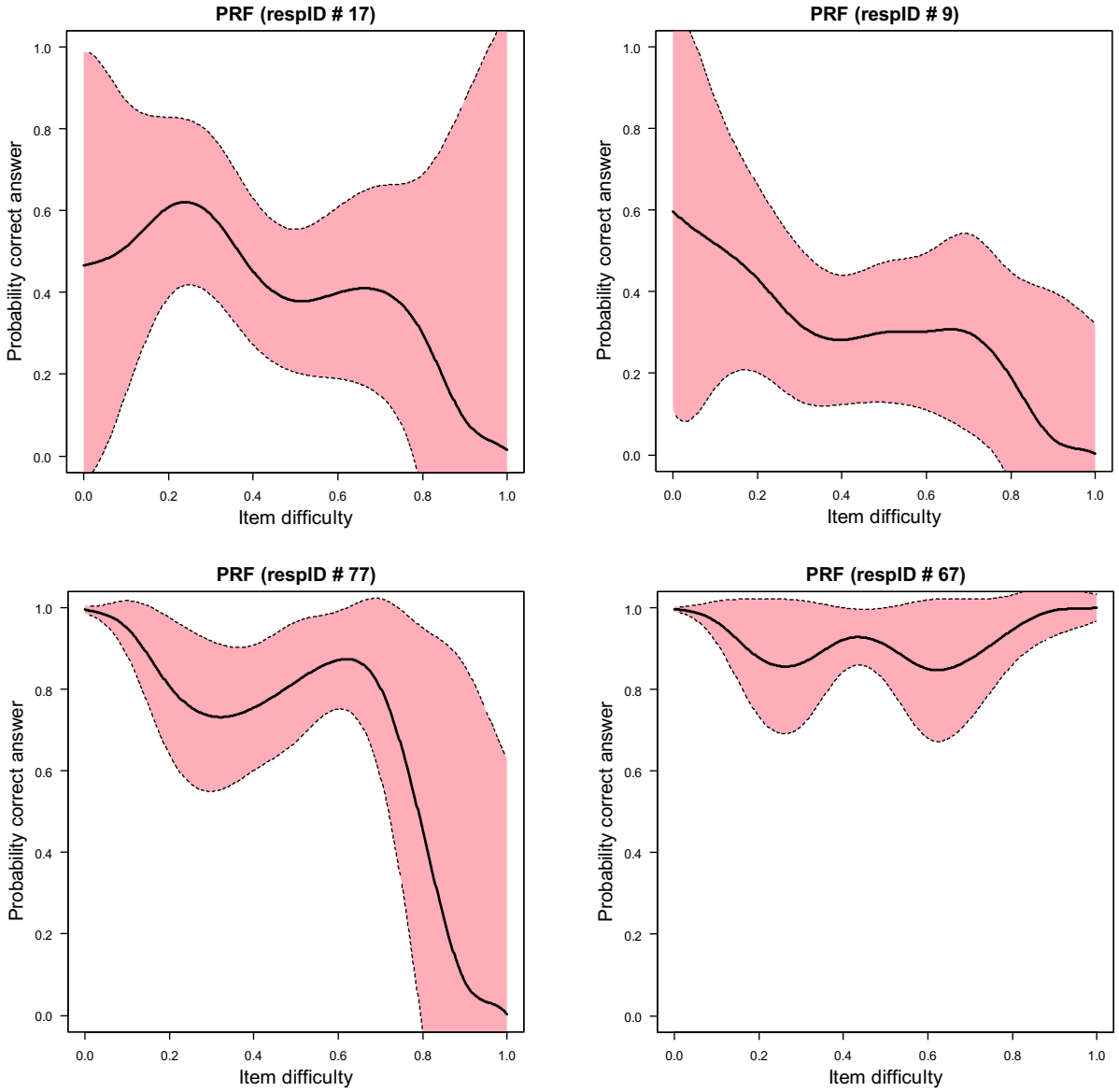


FIGURE 2
Person Response Functions (PRFs) for 10 of the most aberrant responders as identified using the sampling distribution of Ht using bootstrapping. Upper and lower confidence intervals (shaded area) are at 95%.

Cronbach-Mesbach curve. First, an index of person reliability is developed using the K-R 20 formula followed by a person backward stepwise procedure in which one person at a time is deleted from the model. The methodology was applied to the measurement of a quantitative skills latent trait using a sample of 82 participants. Results pointed to the usefulness of the PBRC in identifying aberrant response patterns by flagging 10 participants, who behaved in ways that deviated markedly from the Guttman pattern.

The most important finding of the present study was that the 10 participants flagged using the PBRC were the same 10 worst-fitted participants using the Ht index and were also among the 8 worse participants using the U3 index. Thus, the criterion-related validity of the PBRC was fully supported using Ht and also U3 at a level of 80%. Further, visual analyzes indicated that the PRCs of these participants reflected significant deviations between expected curves and those observed likely being reflective of the processes of lucky guessing (Foley, 2019) and carelessness or inattention (Meade and Craig, 2012; Maniaci and Rogge, 2014). Those participants were across the board of ability with theta values ranging between −1.71 and + 1.79, thus, the methodology was not sensitive to specific levels of person abilities, low or high. The present findings regarding the validity of the Ht and U3 indices corroborated with previous findings showing the superiority of these statistics compared to other alternatives (e.g., Karabatsos, 2003; St-Onge et al., 2011; Rupp, 2013; Tendeiro and Meijer, 2014; Beck et al., 2019; Mousavi et al., 2019; Wongpakaran et al., 2019).

The present study presents visual means to identify aberrant responding and is one of the available tools in data screening so that

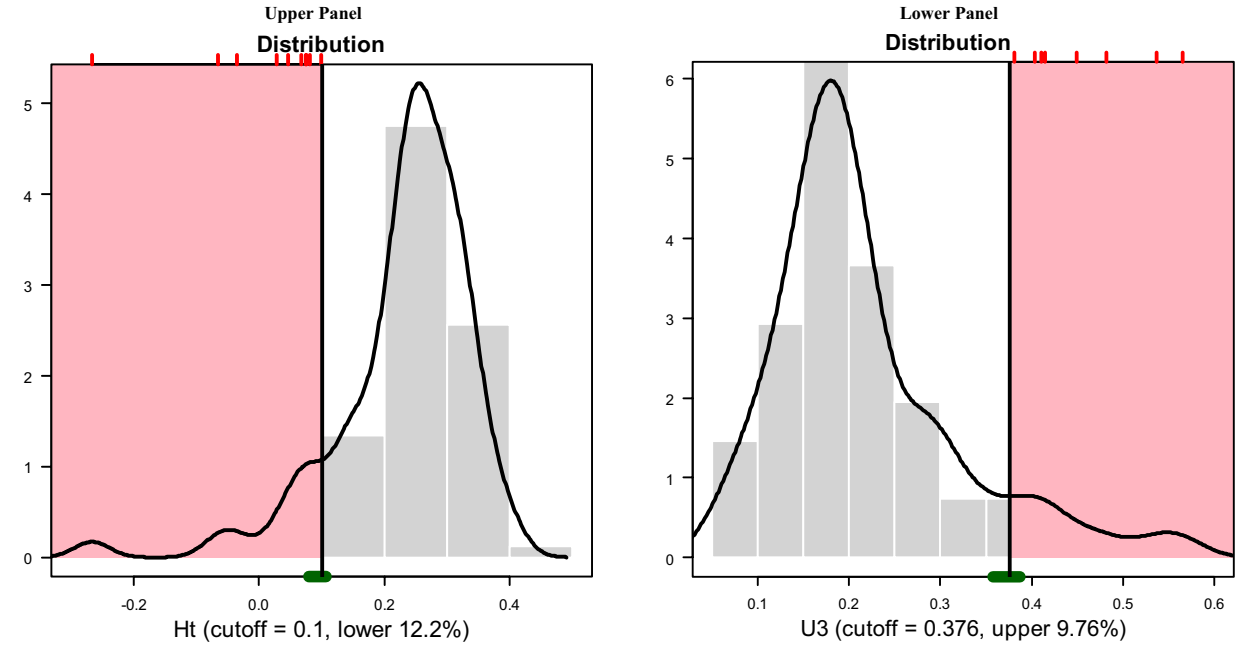


FIGURE 3
Distribution and cutoff value for Ht index.

problematic responders are flagged and potentially removed. Novel ideas beyond person fit indicators involve simulation where response vectors are generated so that they mimic aberrant response patterns. Then these patterns can be evaluated for their presence with real data so that the detection of aberrant responders is achieved (Dupuis et al., 2018).

4.1. Limitations and future directions

The present study is limited for several reasons. First, the sample size was relatively small, and thus, results may have been idiosyncratic. Second, the selection of cutoff values of the person fit indices using bootstrapping represents only one among the different available methodologies (Mousavi et al., 2019). Third, the use of person fit indices is informative only *post hoc*; thus, they cannot inform individuals who may behave in aberrant ways before the study. Not only that but the estimation of person fit indices is based on the estimated item parameters that may also be biased by the presence of misfitting participants. Mousavi et al. (2019) proposed employing an iterative procedure, which may be both complex and cumbersome. Furthermore, as the sample sizes get large, the procedure may become cumbersome in terms of selecting criteria to flag aberrant responders and use criteria based on the level of significance and the expected number of outlying cases using the standard normal.

The currently proposed PBRC will need to be compared to additional aberrant responding indices in the future, such as lz*, and/or other indices that are intended to address particular cases of aberrant response and its underlying processes. The discriminant and predictive validity of the PBRC will need to be assessed in light of the effectiveness of other indicators of aberrant behavior. Future studies may also consider cutoff values and percentage of individuals classified as aberrant responders using both visual and statistical criteria. Additionally, a detailed evaluation of the PBRC's capability and sensitivity to certain sorts of aberrant responses, such as inattention, carelessness, random responding, guessing, and cheating, is required. Researchers may examine the effectiveness of the PBRC in response to particular instances of aberrant behavior by methodically altering these parameters within experimental paradigms. This kind of study may provide crucial validity standards for assessing the PBRC's performance and its capacity to precisely identify and evaluate aberrant responses in various circumstances, populations, and cultures (Van de Vijver and Tanzer, 2004). Researchers may create a framework that might result in the creation of new tools and practices to increase the accuracy and reliability of psychological assessments and educational evaluations by comprehending how PBRC matches with other indices of aberrant behavior (see Bereby-Meyer et al., 2002).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Education and Training Evaluation Commission. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

GS: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. FJ: Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project was funded by ETEC, Riyadh, Saudi Arabia.

Acknowledgments

We would like to acknowledge the support of the Research Department at ETEC in the present study.

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.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1273582/full#supplementary-material

References

Barry, D. (1996). Differential recall bias and spurious associations in case/control studies. *Stat. Med.* 15, 2603–2616. doi: 10.1002/(SICI)1097-0258(19961215)15:23<2603::AID-SIM371>3.0.CO;2-G

Beck, M. F., Albano, A. D., and Smith, W. M. (2019). Person-fit as an index of inattentive responding: a comparison of methods using polytomous survey data. *Appl. Psychol. Meas.* 43, 374–387. doi: 10.1177/0146621618798666

Bereby-Meyer, Y., Meyer, Y., and Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple-choice tests. *J. Behav. Decis. Mak.* 15, 313–327. doi: 10.1002/bdm.417

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., and Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *J. Pers. Soc. Psychol.* 111, 218–229. doi: 10.1037/pspp0000085

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cook, N. E., Faust, D., Meyer, J. F., and Faust, K. A. (2016). The impact of careless and random responding on juvenile forensic assessment: susceptibility of commonly used measures and implications for research and practice. *J. Forensic Psychol. Pract.* 16, 425–447. doi: 10.1080/15228932.2016.1234146

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*. 16, 297–334.

Deribo, T., Kroehne, U., and Goldhammer, F. (2021). Model-based treatment of rapid guessing. *J. Educ. Meas.* 58, 281–303. doi: 10.1111/jedm.12290

Dimitrov, D. M., and Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *J. Appl. Meas.* 7, 170–183.

Drasgow, F., and Parsons, C. K. (1983). Application of unidimensional item response theory models to multitrait-multimethod matrices. *Appl. Psychol. Meas.* 7, 389–416. doi: 10.1177/014662168300700207

Dupuis, M., Meier, E., and Cuneo, F. (2018). Detecting computer-generated random responding in questionnaire-based data: a comparison of seven indices. *Behav. Res. Methods* 51, 2228–2237. doi: 10.3758/s13428-018-1103-y

Emons, W. H., Sijtsma, K., and Meijer, R. R. (2005). Global, local, and graphical person-fit analysis using person-response functions. *Psychol. Methods* 10, 101–119. doi: 10.1037/1082-989X.10.1.101

Ferro, J. M., and Beaton, D. E. (2016). Detecting aberrant responding: a review of methods and applications. *Educ. Meas. Issues Pract.* 35, 18–31.

Foley, B. P. (2019). Getting lucky: how guessing threatens the validity of performance classifications. *Pract. Assess. Res. Eval.* 21:3. Available at: https://scholarworks.umass.edu/pare/vol21/iss1/3

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opin. Q.* 70, 646–675. doi: 10.1093/poq/nfl033

Guttman, L. (1944). A basis for scaling qualitative data. *Am. Sociol. Rev.* 9, 139–150. doi: 10.2307/2086306

Hamon, A., and Mesbah, M. (2002). "Questionnaire reliability under the Rasch model" in *Statistical methods for quality of life studies: Design, measurement and analysis*. eds. M. Mesbah, B. F. Cole and M. L. T. Lee (Boston: Kluwer Academic Publishing), 155–168.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 6, 65–70.

Hong, M., Steedle, J., and Cheng, Y. (2019). Methods of detecting insufficient effort responding: comparisons and practical recommendations. *Educ. Psychol. Meas.* 80, 312–345. doi: 10.1177/0013164419865316

Kam, C. C. S., and Chan, G. H. H. (2018). Examination of the validity of instructed response items in identifying careless respondents. *Personal. Individ. Differ.* 129, 83–87. doi: 10.1016/j.paid.2018.03.022

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Appl. Meas. Educ.* 16, 277–298. doi: 10.1207/S15324818AME1604_2

Leite, W. L., and Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivar. Behav. Res.* 45, 271–293. doi: 10.1080/00273171003680245

Little, T. D., and Moore, K. A. (2013). "Detecting aberrant responding using item response theory" in *The nature of human intelligence*. ed. R. J. Sternberg (New York, NY: Cambridge University Press), 415–439.

Lord, F. M. (1964). The effect of random guessing on test validity. *Educ. Psychol. Meas.* 24, 745–747. doi: 10.1177/001316446402400401

Magis, D., Raiche, G., and Beland, S. (2012). A didactic presentation of Snijders's lz* index of person fit with emphasis on response model selection and ability estimation. *J. Educ. Behav. Stat.* 37, 57–81. doi: 10.3102/1076998610396894

Maniaci, M. R., and Rogge, R. D. (2014). Caring about carelessness: participant inattention and its effects on research. *J. Res. Pers.* 48, 61–83. doi: 10.1016/j.jrp.2013.09.008

McKay, A. S., Garcia, D. M., Clapper, J. P., and Shultz, K. S. (2018). The attentive and the careless: examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. *Comput. Hum. Behav.* 84, 295–303. doi: 10.1016/j.chb.2018.03.007

Meade, A. W., and Craig, S. B. (2012). Identifying careless responses in survey data. *Psychol. Methods* 17, 437–455. doi: 10.1037/a0028085

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Appl. Psychol. Meas.* 18, 311–314. doi: 10.1177/014662169401800402

Meijer, R. R., and Sijtsma, K. (2001). Methodology review: evaluating person-fit. *Appl. Psychol. Meas.* 25, 107–135. doi: 10.1177/01466210122031957

Meisenberg, G., and Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education. *Personal. Individ. Differ.* 44, 1539–1550. doi: 10.1016/j.paid.2008.01.010

Mesbah, M. (2010). "Statistical quality of life" in *Method and applications of statistics in the life and health sciences*. ed. N. Balakrishnan (Hoboken, NJ: Wiley), 839–864.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Meyer, J. F., Faust, K. A., Faust, D., Baker, A. M., and Cook, N. E. (2013). Careless and random responding on clinical and research measures in the addictions: a concerning problem and investigation of their detection. *Int. J. Ment. Heal. Addict.* 11, 292–306. doi: 10.1007/s11469-012-9410-5

Mousavi, A., Cui, Y., and Rogers, T. (2019). An examination of different methods of setting cutoff values in person fit research. *Int. J. Test.* 19, 1–22. doi: 10.1080/15305058.2018.1464010

Müller, J., Hasselbach, P., Loerbroks, A., and Amelang, M. (2015). Person-fit statistics, response sets and survey participation in a population-based cohort study. *Psihologija* 48, 345–360. doi: 10.2298/PSI1504345M

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behav. Ecol.* 15, 1044–1045. doi: 10.1093/beheco/arh107

Paulhus, D. L. (1991). "Measurement and control of response bias" in *Measures of personality and social psychological attitudes*. eds. J. P. Robinson, P. R. Shaver and L. S. Wrightsman (San Diego, CA: Academic Press), 17–59.

Plieninger, H., and Heck, D. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivar. Behav. Res.* 53, 633–654. doi: 10.1080/00273171.2018.1469966

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rios, J. A., Guo, H., Mao, L., and Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: to filter unmotivated examinees or not? *Int. J. Test.* 17, 74–104. doi: 10.1080/15305058.2016.1231193

Rupp, A. A. (2013). A systematic review of the methodology for person-fit research in item response theory: lessons about generalizability of inferences from the design of simulation studies. *Psychol. Test Assess. Model.* 55, 3–38.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika* 66, 331–342. doi: 10.1007/BF02294437

St-Onge, C., Valois, P., Abdous, B., and Germain, S. (2011). Accuracy of person-fit statistics: a Monte Carlo study of the influence of aberrance rates. *Appl. Psychol. Meas.* 35, 419–432. doi: 10.1177/0146621610391777

Tendeiro, J. N., and Meijer, R. R. (2014). Detection of invalid test scores: the usefulness of simple nonparametric statistics. *J. Educ. Meas.* 51, 239–259. doi: 10.1111/jedm.12046

Tendeiro, J. N., Meijer, R. R., and Niessen, A. S. M. (2016). PerfFit: an R package for person-fit analysis in IRT. *J. Stat. Softw.* 74, 1–27. doi: 10.18637/jss.v074.i05

Van de Vijver, F., and Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: an overview. *Eur. Rev. Appl. Psychol.* 54, 119–135. doi: 10.1016/j.erap.2003.12.004

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *J. Cross-Cult. Psychol.* 13, 267–298. doi: 10.1177/0022002182013003001

van Laar, S., and Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: a threat to validity? *J. Educ. Meas.* 59, 470–501. doi: 10.1111/jedm.12317

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., and Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Comput. Hum. Behav.* 76, 417–430. doi: 10.1016/j.chb.2017.06.032

Wise, S. L., and Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Appl. Meas. Educ.* 18, 163–183. doi: 10.1207/s15324818ame1802_2

Wongpakaran, N., Wongpakaran, T., and Kuntawong, P. (2019). Evaluating hierarchical items of the geriatric depression scale through factor analysis and item response theory. *Heliyon* 5:e02300. doi: 10.1016/j.heliyon.2019.e02300

frontiers | Frontiers in Psychology

TYPE Brief Research Report
PUBLISHED 23 October 2023
DOI 10.3389/fpsyg.2023.1268074

OPEN ACCESS

EDITED BY
Hamdollah Ravand,
Vali-E-Asr University of Rafsanjan, Iran

REVIEWED BY
Yi-Hsin Chen,
University of South Florida, United States
Wolfgang Lenhard,
Julius Maximilian University of Würzburg, Germany

*CORRESPONDENCE
Ioannis Tsaousis
ioantsaousis@psych.uoa.gr

RECEIVED 27 July 2023
ACCEPTED 05 October 2023
PUBLISHED 23 October 2023

CITATION
Tsaousis I, Alahmandi MTS and Asiri H (2023) Uncovering Differential Item Functioning effects using MIMIC and mediated MIMIC models.
Front. Psychol. 14:1268074.
doi: 10.3389/fpsyg.2023.1268074

Uncovering Differential Item Functioning effects using MIMIC and mediated MIMIC models

Ioannis Tsaousis[1]*, Maisaa Taleb S. Alahmandi[2] and Halimah Asiri[2]

[1]Department of Psychology, National and Kapodistrian University of Athens (NKUA), Athens, Greece, [2]Education and Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia

The aim of this study was twofold: first, to examine the presence of bias across gender in a scholastic achievement test named the Academic Achievement Test (AAT) for the Science Track. Second, to understand the underlying mechanism that causes these bias effects by examining the effect of general cognitive ability as a mediator. The sample consisted of 1,300 Saudi high school students randomly selected from a larger pool of 173,133 participants to reduce the effects of excessive power. To examine both goals, the Multiple Indicators Multiple Causes (MIMIC) approach for detecting Differential Item Functioning (DIF) items was used. The results showed that 13 AAT items exhibited DIF effects for different gender groups. In most of these items, male participants were more likely to answer them correctly than their female counterparts. Next, the mediated MIMIC approach was applied to explore possible underlying mechanisms that explain these DIF effects. The results from this study showed that general cognitive ability (i.e., General Aptitude Test - GAT) seems to be a factor that could explain why an AAT item exhibits DIF across gender. It was found that GAT scores fully explain the DIF effect in two AAT items (full mediation). In most other cases, GAT helps account for only a proportion of the DIF effect (partial mediation). The results from this study will help experts improve the quality of their instruments by identifying DIF items and deciding how to revise them, considering the mediator's effect on participants' responses.

KEYWORDS

Differential Item Functioning (DIF), uniform DIF, MIMIC approach, mediation analysis, mediated MIMIC model

1. Introduction

In modern psychometrics, there is an increasing interest in identifying and understanding what causes a Differential Item Functioning (DIF) effect (Raykov and Marcoulides, 2011). DIF refers to a situation where an item performs differently across groups of individuals even though those individuals are supposed to have the same level of the trait being measured (Dorans and Holland, 1993). DIF can be caused by cultural, societal, or demographic variables, and it can undermine the fairness and validity of a test or assessment (Ackerman, 1994). DIF can be categorized into two main types: uniform and non-uniform. An item shows uniform DIF when the performance of one group is always superior to another group for each ability level. On the other hand, non-uniform DIF occurs when an item's bias varies across different levels of the latent trait. Therefore, it is important first to identify DIF items and remove them from the scale.

Several statistical methods for identifying items with DIF have been proposed within the Classical Test Theory (CTT) and the Item Response Theory (IRT). Within the IRT framework, the model-based likelihood ratio test is an approach that is typically used to evaluate the significance of observed differences in parameter estimates between groups (Thissen et al., 1993). Other methods include the likelihood ratio goodness-of-fit test (Thissen et al., 1986) and the simultaneous item bias test (SIBTEST) method (Shealy and Stout, 1993). Within the CTT framework, the Mantel–Haenszel (MH) approach (Holland and Thayer, 1988) and the logistic regression (LR) procedure (Swaminathan and Rogers, 1990) are some of the most popular approaches.

Structural Equation Modelling (SEM) also provides a comprehensive framework for examining and understanding the DIF issue (Camilli and Shepard, 1994). Within this context, several different methods have been suggested, including the Multi-Group CFA method (MG-CFA; Pae and Park, 2006), the modification indices method (Chan, 2000), and the Multiple-Indicator, Multiple-Causes approach (MIMIC; MacIntosh and Hashim, 2003). One of the major advantages of the MIMIC approach over the MG-CFA method is that it uses the entire sample of responses to estimate model parameters and test for DIF (Chun et al., 2016). In this case, the total sample size needed for detecting DIF is smaller than that needed in the MG-CFA approach, where model parameters are estimated separately for each contrasted group (Muthén, 1989). Additionally, several explanatory variables (e.g., demographic) can be included within a MIMIC model, allowing us to identify possible causes of DIF. An example of a MIMIC DIF model is shown in Figure 1 (upper panel), in which a grouping variable (Gender) has direct effects on the items of the scale (e.g., AAT_i) and the latent mean (e.g., scholastic achievement) simultaneously.

Recently, Cheng et al. (2016) proposed a method for detecting DIF items in which they combined the MIMIC methodology with mediation analysis to uncover possible causes of DIF effects. In mediation analysis, it is hypothesized that the independent variable (e.g., Gender) affects the dependent variable (e.g., the item AAT_i) *via* an intervening variable called the mediator (e.g., GAT Score) (Baron and Kenny, 1986). The effect of the mediator in the relationship between the independent and dependent variables can be either full (the direct relationship between Gender and AAT_i disappears after the effect of the mediator is controlled) or partial (the mediator can only explain a part of the relationship between the Gender and AAT_i). This relationship constitutes a uniform DIF and is graphically presented in Figure 1 (lower panel).

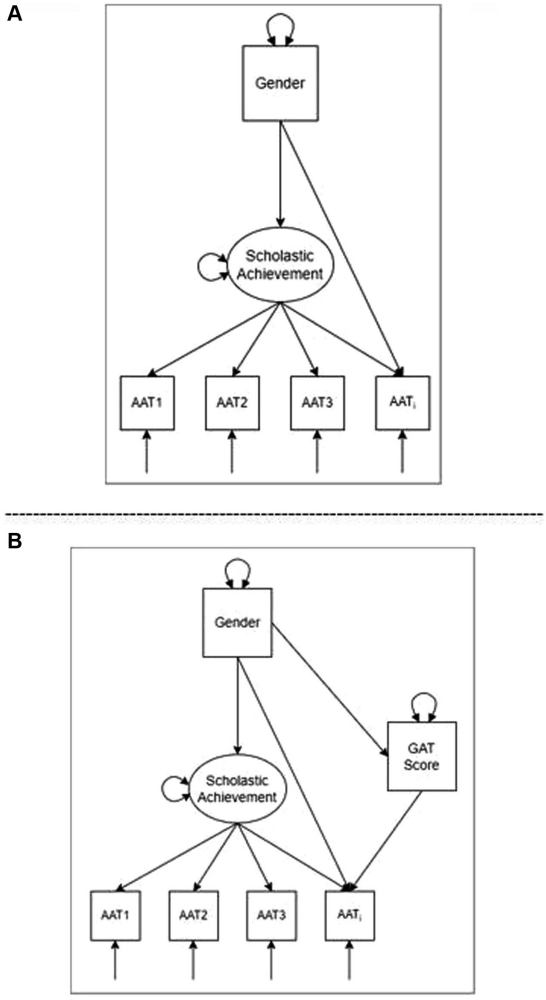


FIGURE 1
MIMIC and mediated MIMIC models for testing DIF effects. **(A)** The standard MIMIC approach to detecting DIF. **(B)** The mediated MIMIC approach to detecting DIF.

2. Research purpose and specific aims

Previous studies have shown that gender is assumed to considerably affect students' academic performance since many studies have shown that boys and girls perform differently (e.g., Voyer and Voyer, 2014). Nevertheless, not all studies agree on the direction and magnitude of this difference (e.g., Else-Quest et al., 2010), and the gender gap in academic attainment is still an open question. This study uses gender as a grouping variable to examine possible DIF effects on academic achievement. It was hypothesized that the response to an AAT item (e.g., AAT_i), which measures scholastic achievement (i.e., the latent variable), involves some general cognitive ability level (i.e., the mediator). Thus, cognitive ability, as measured by the General Aptitude Test (GAT), will completely or partially mediate the relationship between gender and a response to an AAT item when controlling for scholastic achievement. In this study, only uniform DIF was examined.

3. Methods

3.1. Participants and procedure

Previous simulation studies on Differential Item Functioning (DIF) and mediation analysis suggested that with a sample size as large as 1,000 or up and a mediation effect of 0.10 or up, the analysis has enough power to provide robust results (Cheng et al., 2016). Therefore, to reduce the effects of excessive power, a sample of 1,300 participants was randomly selected from a larger pool of 173,133 high school students who completed an achievement test as part of a national examination process. Of them, 648 (49.8%) were males, and 652 (50.2%) were females. The participants' mean age was 17.99 (SD = 0.53). In terms of place of residence, participants originated from all 13 regions of Saudi Arabia. The study was conducted in accordance with the Declaration of Helsinki and approved by the

Institutional Review Board (or Ethics Committee) of the Education & Training Evaluation Commission (Approval Code: TR369-2023, Approval Date: 20/11/2022).

3.2. Measures

3.2.1. The academic achievement test for the science track (AAT; education and training evaluation commission - ETEC)

The AAT is a 44-item admission test that measures achievement level in accordance with university study readiness standards. It consists of four subscales that focus on the general outcomes of the following courses: First, second-, and third-year Biology (12 items), Chemistry (10 items), Physics (10 items), and Mathematics (12 items) of the secondary stage (grades 10, 11, and 12). The AAT test items are in a multiple-choice format and are scored as correct (1) or wrong (0). The test has a 50-min duration and is presented in Arabic.

3.2.2. General aptitude test (GAT) for science major (education and training evaluation commission - ETEC)

This is a general cognitive ability test developed in the Arabic language that measures analytical and deductive skills. It is composed of two cognitive domains: (a) language-related skills (68 items) and (b) numerical-related skills (52 items). Each domain comprises several subdomains, including word meaning, sentence completion, reading comprehension, arithmetic, analysis, geometry, etc. The global cognitive ability factor composed of the scores from the two domain scales was the only available score from this test in this study. All scores were transformed into standard scores (T-scores), with a range of 0–100.

3.3. Data analysis

Before examining DIF effects and possible causes within the Structural Equation Modeling (SEM) framework, the measurement model specification of each of the four AAT scales was examined. The following goodness of fit indices were used: the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMR). CFI and TLI values higher than 0.90 indicate an acceptable fit (with values >0.95 being ideal), and RMSEA and SRMR values up to 0.08 indicate a reasonable fit (with values <0.05 indicating an excellent fit (Hu and Bentler, 1999).

Next, the MIMIC model approach was used to detect DIF items across the different AAT scales. The *MIMIC model with scale purification* (*M-SP*) method was used (Wang and Shih, 2010) for each scale separately. In this approach, the direct effect of the grouping variable (e.g., gender) on an item response (e.g., AAT_i) is estimated. In Figure 1 (upper panel), this relationship is represented by a direct path from Gender to item AAT_i. The direct effect represents the difference in item response between the two levels of the grouping variable (i.e., males vs. females) given the same scholastic achievement ability (latent variable). If the direct effect is significant, this indicates a DIF effect. The indirect effect is represented by a path from grouping variable to latent variable and indicates whether the mean of the latent variable across groups is different. The same procedure will be followed for all AAT items, one at a time. It should also be noted that Bonferroni correction will be adopted to control for the Type I error (Dunn, 1961).

After identifying DIF items, the mediated MIMIC approach was used to uncover possible causes of the emerging DIF effects. As discussed earlier, a mediator (e.g., GAT score) can mediate the relationship between group membership (e.g., gender) and an item response (AAT_i), conditioning on the latent trait (e.g., scholastic achievement). Therefore, when we fit a DIF item (found in the previous analysis step) in the mediation model, we obtain direct and indirect effects for each model. If the direct effect (from the grouping variable to the item) becomes non-significant when the mediator (i.e., GAT score) is taken into account in this relationship (from the grouping variable to the mediator and then to the item), we have full mediation (the indirect effect is significant). This means that the mediator fully explains the DIF effect. On the other hand, if the direct effect is still significant when the mediator is entered into the equation, and the indirect effect is significant, we have partial mediation. In this case, the mediator explains to some extent the DIF effect, but maybe additional mediators are needed to explain the causes of the DIF effect fully. All analyses were conducted using Mplus 8.03 (Muthén and Muthén, 1998-2018).

4. Results

First, the measurement model of each AAT scale (i.e., Biology, Chemistry, Physics, and Mathematics) was examined *via* CFA. A unidimensional structure for each scale was hypothesized. In Table 1, the results from the CFA are reported. The results showed that all measurement models fit the data very well.

Next, a MIMIC approach was applied to detecting uniform DIF items across gender for all AAT scales. During the process of identifying DIF items, every item within each scale was regressed on the grouping variable, with all other items presumed as non-DIF items and serving as the anchor set. In the grouping variable (i.e., gender), males were coded as 0 (the reference group) and females as 1 (the focal

TABLE 1 Model fit indices for AAT scales.

Scales	χ^2	*df*	CFI	TLI	RMSEA (95% CIs)	SRMR
Biology	79.610*	54	0.973	0.967	0.019 (0.009–0.028)	0.038
Chemistry	77.599**	35	0.984	0.980	0.031 (0.021–0.040)	0.043
Physics	111.354**	35	0.924	0.903	0.041 (0.033–0.050)	0.057
Mathematics	84.707**	54	0.985	0.981	0.021 (0.012–0.029)	0.037

χ^2, chi-square goodness of fit statistic; df, degrees of freedom; CFI, Comparative Fit Index; TLI, Tucker Lewis Index; RMSEA, Root Mean Square Error of Approximation; 95% CIs = 95% Confidence Intervals; SRMR, Standardized Root Mean Square Residual. ** Models are significant at $p < 0.001$; * Models are significant at $p < 0.01$.

10.3389/fpsyg.2023.1268074

group). A negative z value indicates that males at the same level of scholastic achievement as females are more likely to respond to the item correctly. To identify potential DIF items, the following equation was applied:

$$Y_{ij} = \lambda_j * \theta_i + \beta_j z_i + e_{ij}$$

Where,

Υ_{ij} = the latent response for item *j* for participant *i*.

λ_j = the factor loading of item *j*.

θ_i = the latent ability of the participant *i*.

z_i = the grouping indicator of the participant *i*.

β_j = the regression coefficient of the corresponding grouping variable, and.

e_{ij} = the random error term.

If β_j is non-significant, then item *j* is the same across groups of variable z_i. However, if β_j is significant, it designates a difference in the response probabilities across groups of variable z_i, designating a DIF item. Practically, DIF is detected when the direct relationship between the group variable (gender) and the item in question is statistically significant. It should be noted that the Benjamini-Hochberg correction was applied to control for false discovery rate (Benjamini and Hochberg, 1995). Table 2 presents the results from the DIF analysis.

The analysis uncovered 13 DIF items. For example, in the Biology scale, items 7 and 8 were detected as DIF items. In item 7, the z value (−2.888) indicates that controlling for scholastic achievement, a male participant is more likely to respond correctly than a female participant. In item 8, on the other hand, the positive z value indicates that female participants are more likely to respond correctly than male participants, although they are at the same level of scholastic achievement.

After this step, the mediated MIMIC approach was applied in an attempt to understand what causes DIF in these items. It was hypothesized that general cognitive ability (i.e., GAT) could be a mediator that mediates the relationship between the grouping variable and the response to a specific item. Table 3 presents the results of the mediation analysis within a MIMIC model.

The results showed that cognitive ability seems to be a factor that could explain why an AAT item exhibits DIF across gender. GAT fully explains the DIF effect in two AAT items (i.e., Chem18 and Chem20) since the direct effect is no longer significant after the mediator enters the equation (full mediation). In both cases, the effect of the GAT score on the probability of correct response is positive ($a_7 = 0.323$, SE = 0.048, z = 6.723, $p = 0.001$, and $a_8 = 0.265$, SE = 0.034, z = 6.074, p = 0.001, respectively). This means that the higher the GAT score, the higher the probability of answering the item correctly. However, the direct effect on both items is negative ($\beta_7 = -0.056$, SE = 0.036, $p = 0.121$, and $\beta_8 = -0.048$, SE = 0.034, $p = 0.155$). This finding suggests that females with the same GAT score are less likely to answer this item correctly compared to males.

In most other cases, GAT helps account for only a proportion of the DIF effect (partial mediation). Obviously, additional factors intervene in the relationship between gender and answering an item correctly and cause DIF effects. Only in one case (i.e., Phys26) could GAT not explain why male participants are more likely to respond correctly to this item than female participants, although both are at the same underlying level of cognitive ability. Interestingly, males were more likely to respond correctly to some items than females (i.e., Bio7, Chem15, Chem18, Chem20, Phys28, and Math34). But when the GAT

TABLE 2 MIMIC examination for DIF across gender.

Items	Estimate (β)	S.E.	*z* value	*p*-value
Biology scale				
Bio1	−0.044	0.036	−1.226	ns
Bio2	−0.008	0.032	−0.255	ns
Bio3	0.042	0.031	1.333	ns
Bio4	0.028	0.032	0.867	ns
Bio5	−0.055	0.032	−1.697	ns
Bio6	−0.013	0.033	−0.405	ns
Bio7	−0.099	0.034	−2.888	0.004
Bio8	0.095	0.031	3.027	0.002
Bio9	0.064	0.032	2.015	ns
Bio10	−0.088	0.037	−2.393	ns
Bio11	0.034	0.032	1.059	ns
Bio12	0.031	0.031	0.974	ns
Chemistry scale				
Chem13	0.116	0.036	3.214	0.001
Chem14	−0.012	0.033	−0.372	ns
Chem15	−0.121	0.037	−3.316	0.001
Chem16	0.050	0.038	1.322	ns
Chem17	0.046	0.031	1.481	ns
Chem18	−0.100	0.034	−2.910	0.004
Chem19	0.065	0.031	2.0101	ns
Chem20	−0.080	0.032	−2.456	0.014
Chem21	−0.022	0.033	−0.668	ns
Chem22	0.023	0.034	0.067	ns
Physics scale				
Phys23	0.056	0.034	1.638	ns
Phys24	0.018	0.032	0.570	ns
Phys25	−0.166	0.040	−4.114	0.001
Phys26	−0.177	0.041	−4.330	0.001
Phys27	−0.083	0.045	−1.845	ns
Phys28	−0.117	0.037	−3.199	0.001
Phys29	0.140	0.032	4.409	0.001
Phys30	−0.048	0.035	−1.371	ns
Phys31	0.186	0.031	5.921	0.001
Phys32	−0.063	0.037	−1.689	ns
Mathematics scale				
Math33	0.023	0.033	0.0700	ns
Math34	−0.128	0.031	−4.143	0.001
Math35	−0.068	0.032	−2.146	ns
Math36	0.077	0.032	2.391	ns
Math37	−0.042	0.032	−1.319	ns
Math38	−0.008	0.032	−0.258	ns
Math39	−0.023	0.032	−0.718	ns
Math40	−0.029	0.032	−0.913	ns
Math41	0.052	0.031	1.706	ns
Math42	0.023	0.041	0.552	ns
Math43	0.085	0.030	2.784	0.005
Math44	0.012	0.033	0.375	ns

Bio, Biology; Chem, Chemistry; Phys, Physics; Math, Mathematics.

TABLE 3 Direct and indirect (mediation) effects for DIF items.

Item	Direct effect	*p*-value	Indirect effect	*p*-value
Bio7	−0.107	0.002	0.019	0.006
Bio8	0.090	0.005	0.022	0.001
Chem 13	0.113	0.003	0.039	0.001
Chem 15	−0.092	0.017	0.026	0.004
Chem 18	−0.056	0.121	0.040	0.001
Chem 20	−0.048	0.155	0.033	0.001
Phys25	−0.166	0.001	0.023	0.003
Phys26	−0.178	0.001	−0.003	0.595
Phys28	−0.114	0.002	0.036	0.001
Phys29	0.142	0.001	0.029	0.001
Phys31	0.190	0.001	0.020	0.001
Math34	−0.136	0.001	0.021	0.002
Math43	0.081	0.010	0.025	0.001

Bio, Biology; Chem, Chemistry; Phys, Physics; Math, Mathematics.

score was taken into account (i.e., as a mediator), the probability of correctly answering these items was higher for females than for males.

5. Discussion

The aim of this study was twofold: first, to examine whether there are gender differences in the probability of correctly answering an item of the AAT. In other words, whether there are DIF items in terms of gender. Second, to understand the underlying mechanism that causes these DIF effects. The first aim, detecting DIF items, was examined *via* a MIMIC approach. MIMIC models have been used extensively for identifying items with DIF (Muthèn, 1985) since it has been found that they work equally well with other methods (Woods, 2009). This study used a MIMIC model to detect possible DIF items across gender for a scholastic achievement test (i.e., AAT). The analysis revealed that 13 AAT items exhibited DIF across gender (i.e., two from the Biology scale, four from the Chemistry scale, five from the Physics scale, and two from the Mathematics scale). Furthermore, in most (9 out of 13), male participants were more likely to answer the items correctly than their female counterparts.

The second aim of this study, to uncover possible causes of DIF, was examined *via* the mediated MIMIC approach. Mediation analysis is a statistical method that provides a framework for understanding why certain phenomena in the relationship among variables occur. Using this analysis within a MIMIC model for detecting DIF, we can explore possible underlying mechanisms that explain these DIF effects. It was hypothesized that general cognitive ability, as measured by the General Aptitude Test (GAT), could mediate the relationship between the grouping variable (e.g., gender) and the response to a specific item. If a mediation effect exists, we can explain why a DIF effect occurs, depending on the Type of mediation (full or partial).

The results from this study showed that general cognitive ability fully explains the DIF effect in two AAT items (i.e., Chem18 and Chem20). In most other cases, GAT helps account for only a proportion of the DIF effect (partial mediation). It seems that additional factors intervene in the relationship between gender and answering an item correctly and cause DIF effects. Interestingly, from all detected DIF items, only for one item (Phys26), GAT could not explain why the DIF effect occurred.

This study offers valuable information regarding DIF effects and the possible causes of these effects. Using the MIMIC approach, DIF effects were examined within the mediation analysis framework. As a result, it was revealed that general cognitive ability mediates the relationship between gender and the probability of success in an item and provides a context for understanding the underlying mechanism of why DIF effects occurred. Therefore, this study will help experts improve the quality of their instruments by identifying DIF items and deciding how to revise them, considering the mediator's effect on participants' responses. Taking the Biology scale as an example, when Subject Matter Experts (SMEs) are asked to generate items, they should pay careful attention to producing items that are purely related to specific knowledge (i.e., physics) rather than general cognitive ability.

The present study also has certain limitations. First, only GAT scores were available as potential mediators. Future studies should explore the role of other variables, including cognitive (e.g., GPA) and emotional (e.g., self-efficacy) constructs, that could be used to explain the emergence of DIF effects. Second, only gender was examined as a potential grouping variable. In future studies, additional variables (e.g., Type of school: public vs. private) could be examined as potential causes of DIF. Finally, in this study, only uniform DIF was investigated. We would like to expand this approach to examine also non-uniform DIF effects. This type of DIF examines whether an item discriminates differently between the groups in question. Thus, important information about non-uniform DIF effects could be revealed by conceptualizing DIF within the context of moderated mediation analysis (Montoya and Jeon, 2020).

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data that supports the findings of this study are available from the Education and Training Evaluation Commission (ETEC). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the authors upon reasonable request and with the permission of the ETEC. Requests to access these datasets should be directed to MA, m.ahmadi@etec.gov.sa.

Author contributions

IT: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. MA: Methodology, Writing – review & editing. HA: Data curation, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was funded by the Education & Training Evaluation Commission (ETEC).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1268074/full#supplementary-material

References

Ackerman, T. A., and Evans, J. A. (1994). The Influence of Conditioning Scores In Performing DIF Analyses. *Applied Psychological Measurement* 18, 329–342.

Baron, R. M., and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51, 1173–1182. doi: 10.1037/0022-3514.51.6.1173

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Camilli, G., and Shepard, L. A. (1994). *Methods for identifying biased test items*. London: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton adaption-innovation inventory using multiple-group mean and covariance structure analyses. *Multivar. Behav. Res.* 35, 169–199. doi: 10.1207/S15327906MBR3502_2

Cheng, Y., Shao, C., and Lathrop, Q. N. (2016). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educ. Psychol. Meas.* 76, 43–63. doi: 10.1177/0013164415576187

Chun, S., Stark, S., Kim, E. S., and Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: exploring a new sequential-free baseline procedure. *Appl. Psychol. Meas.* 40, 486–499. doi: 10.1177/0146621616659738

Dorans, N. J., and Holland, P. W. (1993). "DIF detection and description: mantel-Haenszel and standardization" in *Differential item functioning*. eds. P. W. Holland and H. Wainer (Hillsdale, NJ: Lawrence Erlbaum), 35–66.

Dunn, J. O. (1961). Multiple comparisons among means. *J. Am. Stat. Assoc.* 56, 52–64. doi: 10.1080/01621459.1961.10482090

Else-Quest, N. M., Hyde, J. S., and Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychol. Bull.* 136, 103–127. doi: 10.1037/a0018053

Holland, P. W., and Thayer, D. T. (1988). "Differential item performance and the mantel-Haenszel procedure" in *Test validity*. eds. H. Wainer and H. I. Braun (Hillsdale, NJ: Lawrence Erlbaum), 129–145.

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

MacIntosh, R., and Hashim, S. (2003). Variance estimation for converting MIMIC model parameters to IRT parameters in DIF analysis. *Appl. Psychol. Meas.* 27, 372–379. doi: 10.1177/0146621603256021

Montoya, A. K., and Jeon, M. (2020). MIMIC Models for Uniform and Nonuniform DIF as Moderated Mediation Models. *Applied psychological measurement* 44, 118–136.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika* 54, 557–585.

Muthèn, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *J. Educ. Stat.* 10, 121–132. doi: 10.3102/10769986010002121

Muthén, L. K., and Muthén, B. O. (1998-2018). *Mplus User's Guide. 8th* Edn. Los Angeles, CA: Muthén & Muthén.

Pae, T. I., and Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Lang. Test.* 23, 475–496. doi: 10.1191/0265532206lt338oa

Raykov, T., and Marcoulides, G. A. (2011). *Introduction to psychometric theory* Routledge.

Shealy, R. T., and Stout, W. F. (1993). "An item response theory model for test bias and differential item functioning" in *Differential item functioning*. eds. P. W. Holland and H. Wainer (Hillsdale, NJ: Lawrence Erlbaum Associates), 197–240.

Swaminathan, H., and Rogers, H. J. (1990). Detecting item bias using logistic regression procedures. *J. Educ. Meas.* 27, 361–370. doi: 10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., and Gerrard, M. (1986). Beyond group-mean differences: the concept of item bias. *Psychol. Bull.* 99, 118–128. doi: 10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., and Wainer, H. (1993). "Detection of differential item functioning using the parameters of item response models" in *Differential item functioning*. eds. P. W. Holland and H. Wainer (Hillsdale, NJ: Lawrence Erlbaum Associates), 67–113.

Voyer, D., and Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychological Bulletin* 140, 1174–1204. doi: 10.1037/a0036620

Wang, W. C., and Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Appl. Psychol. Meas.* 34, 166–180. doi: 10.1177/0146621609355279

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivar. Behav. Res.* 44, 1–27. doi: 10.1080/00273170802620121

frontiers | Frontiers in Psychology

TYPE Methods
PUBLISHED 27 December 2023
DOI 10.3389/fpsyg.2023.1298534

OPEN ACCESS

EDITED BY
Ioannis Tsaousis,
National and Kapodistrian University of Athens, Greece

REVIEWED BY
Jung Yeon Park,
George Mason University, United States
Franca Crippa,
University of Milano-Bicocca, Italy

*CORRESPONDENCE
Johnson Li
✉ johnson.li@umanitoba.ca

RECEIVED 21 September 2023
ACCEPTED 08 December 2023
PUBLISHED 27 December 2023

CITATION
Cheng Y, Pérez-Díaz PA, Petrides KV and Li J (2023) Monte Carlo simulation with confusion matrix paradigm – A sample of internal consistency indices.
Front. Psychol. 14:1298534.
doi: 10.3389/fpsyg.2023.1298534

Monte Carlo simulation with confusion matrix paradigm – A sample of internal consistency indices

Yongtian Cheng[1], Pablo A. Pérez-Díaz[2], K. V. Petrides[1] and Johnson Li[3]*

[1]Division of Psychology and Language Sciences, University College London, London, United Kingdom, [2]Instituto de Psicología, Universidad Austral de Chile, Puerto Montt, Chile, [3]Department of Psychology, University of Manitoba, Winnipeg, MB, Canada

Monte Carlo simulation is a common method of providing empirical evidence to verify statistics used in psychological studies. A representative set of conditions should be included in simulation studies. However, several recently published Monte Carlo simulation studies have not included the conditions of the null distribution of the statistic in their evaluations or comparisons of statistics and, therefore, have drawn incorrect conclusions. This present study proposes a design based on a common statistic evaluation procedure in psychology and machine learning, using a confusion matrix with four cells: true positive, true negative, false negative modified, and false positive modified. To illustrate this design, we employ an influential Monte Carlo simulation study by Trizano-Hermosilla and Alvarado (2016), which concluded that the Omega-indexed internal consistency should be preferred over other alternatives. Our results show that Omega can report an acceptable level of internal consistency (i.e., > 0.7) in a population with no relationship between every two items in some conditions, providing novel empirical evidence for comparing internal consistency indices.

KEYWORDS

Monte Carlo simulation, confusion matrix, null distribution, internal consistency, statistic comparison

Introduction

Simulation studies use computer-generated data to investigate research questions (Beaujean, 2018). Monte Carlo simulation is a commonly used procedure that uses repeated random number selections to solve modeling problems (Gelfand and Smith, 1990). It is especially useful when a statistical assumption (e.g., normality) is violated or in situations without theoretical distribution (Fan, 2012). The Monte Carlo simulation was introduced to psychometrics by Patz and Junker (1999a,b).

Psychological researchers are often interested in determining the sampling distributions of test statistics, comparing parameter estimators (e.g., Cohen's *d*), and comparing multiple statistics that perform the same function. In a Monte Carlo simulation context, a key factor is the design of the specific conditions to evaluate.

Different simulation studies use different designs with a variety of conditions. This is because the study's aims usually dictate the selection of the conditions. Suppose a Monte Carlo study has been designed to test the violation of a certain assumption (e.g.,

normality). Both the condition that the assumption has been met (e.g., normal distributed population) and the condition that the assumption has been violated (e.g., skewed distributed population) should be included in the study. Now, assume a different study has been designed to test the performance of a statistic (or several statistics) across different population distributions. In this case, multiple population distributions should be included in the study design. In general, "the major factors that may potentially affect the outcome of interest should be included" (Fan, 2012, p. 437).

However, some recent studies overlooked the inclusion of the null distribution of statistics in the simulations. A null distribution of statistics represents a scenario with no estimated relationship between variables within a given sample (Hunter and May, 2003; Spurrier, 2003).

In this study, we advocate for including a null distribution of statistic conditions in the Monte Carlo simulation when evaluating and comparing statistical measures. Furthermore, we suggest that the performance of a statistic should be assessed in light of commonly used cut-offs. Psychologists often employ informal tests in their research to compare the statistics values to a pre-determined cut-off to reach a binary decision. For example, an area under the curve (AUC) greater than 0.7 in ROC analysis is considered the minimum acceptable threshold (Streiner and Cairney, 2007); A Root Mean Square Error of Approximation (RMSEA) of 0.08 is regarded as the upper limit for Structural Equation Model fitting (SEM, Fabrigar, 1999). An internal consistency (e.g., Cronbach Alpha) greater than 0.7 is considered an acceptable level of reliability, according to Taber (2018). Trizano-Hermosilla and Alvarado (2016) have conducted a Monte Carlo simulation study with a focus on internal consistency performance evaluation. In this paper, we will utilize the influential study by Trizano-Hermosilla and Alvarado (2016) as a practical example to demonstrate the inclusion of a null distribution and the assessment of the statistic using commonly used cut-offs.

This paper is organized into several sections. We review current practices regarding including null distribution in psychological Monte Carlo simulation studies and their associated limitations. Subsequently, we introduce a simulation design rooted in the confusion matrix as a proposed solution. The study conducted by Trizano-Hermosilla and Alvarado (2016) will be used as a practical example of this design. In conclusion, we engage in a comprehensive discussion about the design, supplemented by another illustrative sample.

The null distributions conditions included in the Monte Carlo simulation psychological studies

We observed that the null distribution of statistics is generally included in existing Monte Carlo simulation studies in two ways: First, the null distribution of statistics is included to represent the condition that there is no true mean difference between two groups of scores and are usually referred to as conditions of null effect (e.g., Derrick et al., 2016; Carter et al., 2019; Fernández-Castilla et al., 2021). This is consistent with the suggestion of the American Psychological Association (APA) guidelines. That is, researchers should include the null distribution of statistics (i.e., no mean difference between two groups; Fan, 2012) in any simulation of effect to test and evaluate the potential threat of Type I error.

Second, the researchers include the condition of a null distribution in factors in the simulation (e.g., Heggestad et al., 2015). In Greene et al.'s (2019) study evaluating the bias of different kinds of fit indices, the authors manipulated (a) the strength of the cross-loadings between factors as 0, 0.1, 0.3, and 0.5, (b) the strength of the between-factor correlated residuals as 0, 0.1, 0.3, and 0.5, and (c) the strength of the within-factor correlated residuals as 0, 0.1, 0.3, and 0.5 in a model. In this sample, 0 represents the condition in which the relationship of cross-loadings or correlated residuals does not exist in the population of variables.

In summary, researchers commonly include the null distribution of the statistic condition when estimating a statistic's performance closely related to the mean difference. For example, when examining Cohen's d in a Monte Carlo simulation study, researchers typically include a condition of no mean difference between two populations. Researchers also include the conditions of null distribution in factors in simulation studies for statistical comparison. However, psychological researchers may sometimes neglect to include the null distribution of the statistic in some other circumstances, such as in cases where the examined statistic does not have a close relationship with the mean difference.

Returning to the fit indices study (Greene et al., 2019) one paragraph above, the authors should include conditions that a null distribution in factors, such as no between-factor correlated residual, and the conditions with the null distribution of the statistic, such that some simulated samples should have no relationship with the proposed model (i.e., no model fitting). In our view, the failure to include conditions of null distribution weakens the conclusion of the simulation in the study. This may occur because some researchers have not considered the performance of the statistic in the condition that the dataset follows a null distribution of this statistic. (i.e., how will the fitting index perform on random data?), although other researchers recognize its importance. For instance, Stone (2000, p. 64) points out: "In order to test statistically the fit of an item, it is then necessary to compare the statistic that is calculated with a null distribution." Stone conducted a Monte Carlo simulation based on null distribution to compare goodness-of-fit test statistics in item response theory (IRT) models, and the results showed the superiority of the statistic he proposed. Fan and Sivo (2007) and Fisk et al. (2023) examined the performance of fit indices in structural equation modeling (SEM) under conditions of model misspecification. This misspecification refers to discrepancies between the theoretical structure of the model and the simulated dataset.

In summary, the null distribution of the statistic is widely included in NHST-related statistics. Yet, when evaluating a statistic that does not have a close relationship in NHST (e.g., fit indices), psychological researchers sometimes neglect the null distribution condition. This study demonstrates the importance of this issue using the example of an influential simulation study about the several common statistics of internal consistencies and will propose a new design based on a confusion matrix that always includes a test with null distribution in statistics and evaluates the statistics from these conditions. As our example, we have selected a study conducted by Trizano-Hermosilla and Alvarado (2016), which we will henceforth refer to as the "original study" for convenience.

How will the missing null distribution of statistic conditions influence the result of a simulation study?

In the following section, we will offer a general overview of the original study. We will specifically address the shortcomings of not including null distribution of the statistics conditions in their simulation design and propose enhancements through the methodology developed in this study.

In the original study, Trizano-Hermosilla and Alvarado (2016) compared the performances of four internal consistency statistics: Cronbach's Alpha, Omega (McDonald, 1999), GLB (Greatest Lower Bound, Sijtsma, 2009), and GLBa (Greatest Lower Bound algebraic, Moltner and Revelle, 2015). They made a comparison of these statistics with various normal and nonnormal distributions and two kinds of inter-correlation between items: tau-equivalent and congeneric.

The original study used Root mean square error (RMSE) and %bias as their criteria.

$$RMSE = \sqrt{\frac{\Sigma\left(\hat{p} - p\right)^2}{Nr}} \tag{1}$$

$$\%bias = \frac{\Sigma\left(\hat{p} - p\right)}{Nr} \times 100\% \tag{2}$$

where $\hat{p}$ refers to the observed statistics for each replication, p refers to the true value of statistics in the simulation population, and Nr refers to the number of replications. Larger absolute values in the RMSE and the %bias statistics indicate worse performance.

Based on the RMSE and the %bias, the authors reported that Omega showed the best performance across most conditions included in their study. In other words, when comparing the difference between observed sample statistics and the associated true population parameter values, Omega showed the smallest discrepancies across most of the conditions. This led the authors to conclude that Omega should be recommended as the preferred index of internal consistency in psychological research. Specifically, the original study suggests that Omega should be preferred over Cronbach's Alpha, which is the most widely used measure of internal consistency. Various studies across multiple disciplines shared the opinion with the original study that Omega rather than Alpha should be used as an internal consistency measurement method (Watkins, 2017; McNeish, 2018; Cortina et al., 2020).

Importantly, for our purposes, Trizano-Hermosilla and Alvarado (2016) original study included only simulation conditions in which there was an effect measured by the statistic (i.e., populations with internal consistency). Specifically, it only included conditions with an acceptable level of internal consistency between items in the questionnaires (i.e., a true internal consistency of 0.731 and 0.845) for the condition of 6 and 12 questionnaire lengths, respectively. As mentioned above, Alpha and Omega values of 0.7 or above are indicated as acceptable internal consistency in psychological research (Taber, 2018). Therefore, it included the null effect of some factors (e.g., no distribution error). However, it did not include a null distribution statistic condition, as we suggest here. According to Tang et al. (2014), internal consistency refers to the degree of inter-item correlations among items with factor saturation. Thus, to simulate a null distribution for these internal consistency statistics, one can independently assign random numbers to each item.

As a result, we would argue that the conditions included in the original study are insufficient to support their conclusions. To elaborate, we propose a new hypothetical index, C, which is used to measure internal consistency, with 0.7 being set as an acceptable cut-off. C is a constant number that can be computed and observed across all the 1,000 simulated datasets. Suppose C is found to be 0.78 from each replicated sample, i.e., (3)

$$C = 0.78. \tag{3}$$

In other words, C is a dummy index without validity according to internal consistency estimation. However, based on the criteria employed in previous studies (i.e., RMSE and %bias), C has a similar level of error as the Omega index. Across conditions in the original study for length = 6 items, the population parameter of internal consistency is 0.731. This is based on Equations (1) and (2), in which p is always 0.731, and $\hat{p}$ is always 0.78. As a result, the RMSE is 0.049 (4)

$$\text{RMSE} = \sqrt{\frac{\sum_{N_r=1}^{1,000}(0.731 - 0.78)^2}{1000}} = 0.049 \tag{4}$$

and the %bias is −4.9% (5)

$$\%\text{bias} = \frac{\sum_{N_r=1}^{1,000}(0.731 - 0.78)}{1000} \times 100\% = -4.9\% \tag{5}$$

in all conditions. Across conditions included in the original study length of 12 items, with similar calculations, RMSE is 0.065, and the %bias is 6.5%. These two results will remain consistent regardless of other factors, like the type of distribution. Therefore, it appears that in a number of conditions, this dummy index can provide similar or even superior performance to the genuine indices included in the original study. Importantly, this indicates that based only on the empirical evidence provided in the original study, we cannot distinguish Omega from this dummy index C. C is an extreme theoretical case, and a statistic with a consistent number cannot be applied. However, A dummy index similar to C with variations can be simulated easily. For example, $\dot{C}$ can be simulated from a continuous uniform distribution [0.711, 0.751] and $\dot{C}$ also cannot be distinguished from Omega with the simulation conditions and criteria used in the original study.

To sum up, simulation studies often evaluate the performance of a statistic based on RMSE and %bias in Monte Carlo simulations, with a view to quantifying the distance between the sample estimates of an observed statistic and the true parameter values (i.e., TP) in the population. We agree that this approach can offer insights regarding the degree to which observed sample estimates are different from true population values. However, without the introduction of the null distribution of statistic conditions in simulation, researchers may reach incorrect or incomplete conclusions, as in the above example with the dummy C index.

To address this problem, we introduce in this study a Monte Carlo design based on criteria commonly used in psychology and machine learning to evaluate models with categorical or binary results: the "confusion matrix" (Marom et al., 2010). In psychology, researchers typically use a confusion matrix to evaluate the performance of a categorization model in real psychological practice (i.e., Ruuska et al., 2018). A confusion matrix comprises four quadrants: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Their relationships are shown in Table 1.

We use the original study as an example to illustrate how to apply confusion matrix methodology to simulation studies in psychology, aiming at statistical comparison. The null distribution of the statistic in the original study is the condition where there is zero internal consistency between items in the population.

We also include the interpretation of results that come from the design. Internal consistency is both continuous and binary (i.e., cut-off), exemplifying the problem adequately. The present study will keep the original study's design unless alternative designs better target the research problem, or the original design is not applicable.

Simulation Study 1: estimating the true negative

As noted above, the original study does not include the null distribution of the internal consistency statistics. The original study only provides empirical evidence of TP. The following simulation aims to distinguish Omega from a dummy index like *C*. As suggested by several studies (Moriasi et al., 2007; Wang and Lu, 2018), for the continuous variable, RMSE and %bias can provide more information than a percentage. Therefore, we propose TN can also use these two criteria. In the case that Omega is an efficacy statistic or index and that TN conditions should have an RMSE close to 0 and a %bias close to 0%, upon which a dummy statistic or index like *C* should have an RMSE close to 0.70 and a %bias close to 70% in the TN condition. Therefore, using TN to test whether a statistic or index is merely a "dummy" one is crucial, and its inclusion in the simulation represents an important step toward obtaining truly conclusive results.

Design

In the original study, the researchers simulated several factors, including sample size (250, 500, and 1,000) and item length (6 or 12).

TABLE 1 The elements of a confusion matrix.

	Predicted true result	Predicted false result
Actual true result	TP	FN
Actual false result	FP	TN

True Positive (TP) is the proportion of results that correctly indicates the presence of a condition or characteristic in the population; True Negative (TN) is the proportion of results that correctly indicates the absence of a condition or characteristic in the population; False Positive (FP) is the proportion of results that erroneously indicates that a particular condition or attribute is present in the population, while False Negative (FN) is the proportion of results which erroneously indicates that a particular condition or attribute is absent in the population. In an ideal perfect model, TP and TN should be at 100%, and FN and FP should be at 0%.

This study will also use the same design for these two factors. The original study also included the distribution of errors following Headrick's (2002) that were introduced to 2, 4, or all 6 items of the 6-item condition and to 2, 4, 6, 8, 10, or all 12 items of the 12-item condition. However, Headrick's method (2002) was not introduced to our current study to ensure there is no internal consistency created from this method between items and results and also for simplicity.

The original study included the tau-equivalent and congeneric models as simulation conditions. This aspect of the study does not apply to the condition of null distribution to internal consistency statistics. This is because there is no correlation between any two items in the null distribution population, regardless of its type. Therefore, this design is not included in our study. In summary, 2*3 = 6 conditions are included in the first simulation.

The original study simulated all datasets in R (R Core Team, 2021) with R Studio (Racine, 2012). The present study will also use these platforms (for details, please see Appendix). For each condition, the design of the original study was replicated 10,000 times. The current study will use the same replication time with 10,000 across six conditions.

Four kinds of internal consistency measurement indices were included in the original study: Alpha, Omega, GLB, and GLBa. As provided by the sample code in the original study, these functions were used in the original study to obtain the following results: omega$alpha for Alpha, omega$omega.tot for Omega, glb.fa$glb for GLB, and glb.algebraic$glb for GLBa. In addition, two packages were used in calculations in the original studies: Psych (Revelle, 2015) and GPArotation (Bernaards and Jennrich, 2015), which are also used in this study. Further, the Omega.total was used as the chosen index from Omega in the original study because Trizano-Hermosilla and Alvarado also reported and evaluated the performance of ω_t, and consequently, the present study will also make use of the same reliability index.

To create a null distribution of internal consistency, we simulated the dataset from a standard normal distribution N(0,1) for each item across participants during the replications. Accordingly, each item and each participant's response are totally independent, which ensures that the true covariances and factor loadings in the population are always zero. To check the validity of this design, we followed Fan (2012, p. 436), who suggests, "We may do a quick data generation verification by generating a large sample." We simulated a large dataset from N(0,1) and calculated four internal consistency indices, as they yielded results close to zero, which supported the simulated null distribution of statistics as accurate. This part of the code is provided separately. This study will also use RMSE and %bias as criteria, similar to the original study, to evaluate the performance of the statistics.

Results

Our results indicated that none of these indices performed as a dummy index. However, according to the criteria used in the original study, Omega (i.e., Omega.total) performed the worst in some TN conditions and never the best. In contrast, Alpha showed the best performance across all conditions. This is possibly because Omega, by definition, cannot be smaller than zero, implying that errors can only inflate its results. The full results of our simulation are displayed in Table 2.

TABLE 2 Estimation of true negative in Study 1.

		%bias				RMSE			
Length	SS	Alpha	Omega	GLB	GLBa	Alpha	Omega	GLB	GLBa
6	250	12.41%	38.39%	15.89%	30.67%	14.17%	39.35%	17.58%	32.53%
6	500	9.00%	33.42%	11.52%	27.67%	10.35%	34.70%	12.87%	29.73%
6	1,000	6.42%	28.99%	8.23%	24.77%	7.42%	30.63%	9.22%	27.11%
12	250	17.67%	29.27%	26.59%	34.45%	18.62%	29.71%	27.47%	35.52%
12	500	12.76%	23.38%	19.13%	30.34%	13.58%	23.87%	19.87%	31.45%
12	1,000	9.28%	19.15%	13.78%	27.04%	9.90%	19.74%	14.35%	28.26%

Length is the length of items; SS is the sample size, and RMSE is Root Mean Square Error without the degree of freedom adjustment.

It should be noted that this design is subject to a limitation. The performances of statistics, which are RMSE and %bias used in the original study, have different meanings when the true effects are different. For example, a 10%bias in a condition where the true effect is 0.731 can lead to a considerable number of studies establishing wrong predictions that view the internal consistency of a study as not acceptable because it could create a 95% distribution like [0.5, 0.9] and consequently yield wrong decisions based on these outputs. For instance, a researcher could consider a 0.6 measurement error of Alpha level in a study not acceptable. A 20%bias in the condition that the true effect is zero will not usually influence the decision-making process since it could create a 95% distribution like [0, 0.3]. In the scenario of a null distribution of internal consistency in the population, it does not matter whether the internal consistency is 0.1 or 0.2, as neither internal consistency score is acceptable in psychological studies.

We propose adding two new designs to the Monte Carlo simulation in psychological statistic testing to overcome this limitation: FNm and FPm.

Simulation Study 2: estimating the modified false positive and modified false negative

First, it is essential to review the definition of FP and FN in the confusion matrix. As shown in Table 1, an accurate definition of FP is the percentage of results that erroneously indicate that a particular condition or attribute (e.g., correlation between variables in the same test) is present, whereas FN is the percentage of results that erroneously indicate that a particular condition or attribute is absent.

These two percentages can be used as criteria in binary outcomes. However, their usage with continuous results is problematic. FP and FN are originally designed for binary results (e.g., yes or no, acceptable or unacceptable). In computer science, the results tend to be clear and objective (i.e., an object is a dog or not a dog). However, this is not always the case in psychological science. The pre-determined cut-off used in psychology for binary conclusions is arbitrary. For instance, it is hard to justify why 0.69 is an unacceptable level of internal consistency while 0.70 is acceptable. This kind of binary thinking is often inappropriate in psychological research. It further implies that designing and measuring P (Internal consistency in simulation <0.70| Population internal consistency = 0.71) or P (Internal consistency in simulation >0.70| Population internal consistency = 0.69) becomes questionable since there is no substantive difference between an internal consistency of 0.69 and 0.70, in which P (X|Y) is a conditional probability, means the possibility of X in the condition of Y.

In addition, as discussed above, it is also meaningless to measure the percentage of internal consistency and report a weak relationship among variables when the internal consistency in the population follows a null distribution (i.e., P (Internal consistency in simulation >= 0.05| Population internal consistency = 0) and (Internal consistency in simulation <= 0.05| Population internal consistency = 0)). Thus, internal consistency close to 0.1 is not acceptable in psychological research.

Therefore, these FP and FN percentages have little practical meaning. However, there is a clear difference between the null distribution condition (e.g., Internal consistency = 0.0) and an acceptable level of relationship (e.g., Internal consistency = 0.7). Therefore, we propose two new metrics based on FP and FN, named FPm and FNm, and suggest a study similar to the original that additionally measures these metrics, in which FPm is the percentage that a statistic returns an acceptable level of statistics result when the statistic follows a null distribution in the population (6), and FNm is the percentage that a statistic returns a null result statistic when, in fact, the parameter is at an acceptable level in the population (7).

$$\mathrm{FPm} = \mathrm{P}\left(\begin{array}{l}\text{Acceptable level of a statistic in}\\ \text{simulation} \mid \text{Null distribution of statistic}\end{array}\right). \quad (6)$$

$$\mathrm{FNm} = \mathrm{P}\left(\begin{array}{l}\text{Null distribution of statistic}\\ \text{in simulation} \mid \text{Acceptable level}\\ \text{parameter in population}\end{array}\right). \quad (7)$$

The letter M in FPm and FNm stands for modification.

According to the percentage that should be measured, the FPm in this study is (8).

$$\mathrm{P}\left(\begin{array}{l}\text{Internal consistency in simulation} >= 0.7 \mid\\ \text{Population internal consistency parameter} = 0.0\end{array}\right). \quad (8)$$

The FNm in this study is (9).

$$\mathrm{P}\left(\begin{array}{l}\text{Internal consistency in simulation} <= 0.0 \mid\\ \text{Population internal consistency} = 0.70\end{array}\right) \quad (9)$$

RMSE is not applicable for this design. Yet, criteria are needed for the purpose of this new design. Hence, we propose two criteria:

Ideally, FPm and FNm should be close to 0% across all conditions. Therefore, for comparison between statistics, the fewer the number of conditions having a number larger than zero, the better the statistic.

In addition, suppose that FPm or FNm is larger than 5% in a certain condition. We suggest that the statistic should be deemed questionable in this condition and not used. This suggestion is based on the standard tolerable level of binary decision error. For instance, if a statistic shows an FPm of 0.3 when the sample size is 200, we would propose that this statistic is unreliable in this sample size condition because an acceptable level of relationship can be reported by this statistic even if the statistic in the population follows a null distribution. However, this statistic could be reliable with a sample size of 1,000, depending on TP, TN, FPm, and FNm values following this rationale. As a result, we suggest that extreme conditions in psychological research should be included in the simulation study to provide comprehensive results.

Design

At first, for both FPm and FNm, the following conditions were included in our study as the original study did: the four internal consistency indices and questionnaire lengths of 6 and 12 items. We included 250, 500, and 1,000 for sample size. In addition, small sample sizes of 20, 25, 30, 35, 40, 45, and 50 are included in this study to test whether there is any condition in psychological studies that these biases will influence TN results.

In the evaluation of FPm, the datasets were simulated with the same population [i.e., N (0.1)] as in Study 1 to create the null distribution of statistics. In the evaluation of FNm, the datasets were simulated with the same method implemented in the original study. This makes the overall conditions 7*2 = 14. Both tau-equivalent and congeneric models are included. The population covariance matrixes are displayed in the code. All four statistics in the original study are included with questionnaire lengths of 6 and 12. Consequently, this makes the overall conditions 14 in FPm and 28 in FNm.

Results

The simulation results of FPm are displayed in Table 3, while the results of FNm are displayed in Table 4. As can be seen in Table 3, *based* on the criteria we proposed, (1) Alpha performs best when there is a null distribution in the internal consistency, and (2) the acceptable level of results of Omega, GLB, and GLBa is questionable when the sample size is less than 30 to 40, depending on the questionnaire length. As can be seen in Table 4, based on the criteria we proposed, all internal consistency indices showed good FNm. This suggests that, under the conditions of our study using the four indices, a result close to zero is highly unlikely to originate from a population with an acceptable level of internal consistency.

Discussion

Our study, alongside the original study by Trizano-Hermosilla and Alvarado (2016), presents a new Monte Carlo simulation design within the confusion matrix paradigm. We have proposed new conditions, guided by the perspective of the confusion matrix, that should be included in the evaluation of statistical simulation studies. Firstly, we will discuss the findings of internal consistency indices. Secondly, we will provide a summary of how to apply this novel confusion matrix design to simulation studies in statistics comparison. Thirdly, we will engage in a general discussion.

TABLE 3 Estimation of false positive modified in Study 2.

Length	SS	Alpha	Omega	GLB	GLBa
6	20	0.19%	**49.93%**	**11.07%**	**8.48%**
6	30	0.00%	**16.55%**	1.73%	1.85%
6	40	0.00%	**5.43%**	0.22%	0.36%
6	50	0.00%	1.63%	0.07%	0.16%
6	250	0.00%	0.00%	0.00%	0.00%
6	500	0.00%	0.00%	0.00%	0.00%
6	1,000	0.00%	0.00%	0.00%	0.00%
12	20	0.89%	**40.98%**	**85.25%**	**35.37%**
12	30	0.00%	**5.03%**	**42.15%**	**11.59%**
12	40	0.00%	0.32%	**14.03%**	3.36%
12	50	0.00%	0.02%	3.70%	0.98%
12	250	0.00%	0.00%	0.00%	0.00%
12	500	0.00%	0.00%	0.00%	0.00%
12	1,000	0.00%	0.00%	0.00%	0.00%

Length is the length of items; SS is the sample size. The percentage values are acceptable (i.e., adequate reliability) when the dataset follows a null distribution (i.e., zero reliability) in the population. Percentages in bold are the Percentages above 5%, which suggests the result of a specific statistic is questionable in this condition.

Issues of internal consistency indices

This study is not primarily focused on which kind of internal consistency indices should be used in psychological research. Therefore, the study has replicated the design of the original study (i.e., sample size and questionnaire length) when applicable to provide an example of how to apply this confusion matrix design. This does not imply that we see no space for improvement in the conditions included in the study. For instance, Likert scale variables should be included in the simulation as internal consistency indexes are usually applied to the Likert scale variables in psychological research (Croasmun and Ostrom, 2011). However, we have found additional empirical evidence that should be used as a reference for the performance of these statistics. Through this additional evidence, we have found that Omega and GLB indices do not perform well enough for small sample sizes under some conditions. Yet, our results do not imply that Alpha should necessarily be preferred over Omega. We admit that Alpha has shortcomings as an index for measuring internal consistency, which is boosted by the length of the questionnaire or prerequisites that are violated, as described in previous studies (McNeish, 2018; Hayes and Counts, 2020).

However, we have found that under some conditions (e.g., sample size = 20, 30, or 40), Omega.total and GLB are boosted and thus become unreliable. Specifically, it is difficult to distinguish a population with random numbers from a population that has high internal consistency. Therefore, in these conditions (i.e., sample size

TABLE 4 Estimation of false negative modified in Study 2.

QL	SS	Condition	Alpha	Omega	GLB	GLBa
6	20	TE	0.28%	0.00%	0.00%	0.00%
6	20	CG	0.34%	0.00%	0.00%	0.00%
12	20	TE	0.01%	0.00%	0.00%	0.00%
12	20	CG	0.01%	0.00%	0.00%	0.00%
6	30	TE	0.09%	0.00%	0.00%	0.00%
6	30	CG	0.15%	0.00%	0.00%	0.00%
12	30	TE	0.03%	0.00%	0.00%	0.00%
12	30	CG	0.01%	0.00%	0.00%	0.00%
6	40	TE	0.03%	0.00%	0.00%	0.00%
6	40	CG	0.03%	0.00%	0.00%	0.00%
12	40	TE	0.01%	0.00%	0.00%	0.00%
12	40	CG	0.00%	0.00%	0.00%	0.00%
6	50	TE	0.00%	0.00%	0.00%	0.00%
6	50	CG	0.03%	0.00%	0.00%	0.00%
12	50	TE	0.02%	0.00%	0.00%	0.00%
12	50	CG	0.02%	0.00%	0.00%	0.00%
6	250	TE	0.00%	0.00%	0.00%	0.00%
6	250	CG	0.01%	0.00%	0.00%	0.00%
12	250	TE	0.00%	0.00%	0.00%	0.00%
12	250	CG	0.00%	0.00%	0.00%	0.00%
6	500	TE	0.00%	0.00%	0.00%	0.00%
6	500	CG	0.01%	0.00%	0.00%	0.00%
12	500	TE	0.00%	0.00%	0.00%	0.00%
12	500	CG	0.00%	0.00%	0.00%	0.00%
6	1,000	TE	0.00%	0.00%	0.00%	0.00%
6	1,000	CG	0.06%	0.00%	0.00%	0.00%
12	1,000	TE	0.00%	0.00%	0.00%	0.00%
12	1,000	CG	0.00%	0.00%	0.00%	0.00%

QL is the length of items; SS is the sample size. TE is tau-equivalent model. CG is Congeneric model. Percentage values are failures that suggest statistics report that there is no internal consistency when in fact, there is an acceptable internal consistency in the population.

<40), we recommend that Omega.total and GLB be avoided in estimating the internal consistency, no matter what kind of performance Omega.total has when there is an acceptable level of the parameters in a given population. These suggestions are based on the results of this simulation study, which are limited by the study's design.

We simulated a null distribution for internal consistency, specifically using a normal distribution generated randomly for each item. This implies that all effects in the dataset are essentially noise. To our understanding, the reason why the Omega statistic tends to be inflated in small sample sizes is due to its value range being restricted to [0,1]. Consequently, any noise in the dataset disproportionately affects Omega positively. As suggested by Okada (2017), the zero-winsorized method can create positive biases. Especially in conditions of small sample sizes, such biases can lead to inflated results, sometimes even exceeding the established cut-off (i.e., 0.7).

Moreover, related Omega indices, such as Hierarchical Omega, should also be tested when researchers aim to measure the reliability of the general factor only. All these indices with these conditions should be tested through the TP and TN conditions, corresponding to FPm and FNm. Most importantly, all these conditions should be tested simultaneously in a simulation study to provide empirical evidence for applied researchers. Suppose the proposed design had been applied in the original study. A more conservative recommendation of Omega with a discussion of Omega's limitations will be provided in the original study and studies influenced by the original study (Watkins, 2017; McNeish, 2018; Cortina et al., 2020).

Practical recommendations and steps when implementing a confusion matrix design through Monte Carlo simulation

Step 1: Both conditions in which there is a certain relationship between variables and the condition in which the expected association is deemed as absent should be included in the simulation design (i.e., the null distribution of statistics), together with other relevant criteria such as sample size, distribution, and alike. These two kinds of conditions ought to be included as TP and TN, respectively.

In simulating the null distribution of statistics, we advocate for consistently employing the method outlined in the APA guidelines (Fan, 2012). This approach ensures that the simulation design accurately represents a population with a null statistic distribution and assesses its impact on the observed sample statistics. Our findings confirm that it is possible to reconstruct an estimation by a normally distributed dataset in the absence of internal consistency across four reliability statistics, which have theoretical and practical implications that are related to the definition and calculation of what is considered to be a large sample. For instance, as described in Study 1, we calculated all four statistics (i.e., Omega, Alpha, GLB, and GLBa) with a large sample of 100,000 and a standard normal distribution, ensuring the inclusion of a null distribution of statistics since all the statistics are close to zero in such an extensive sample.

Meanwhile, it's important to acknowledge that there are various types of null distributions for a statistic. Although our simulation study only includes normal distributions, we encourage researchers to explore a broader range of nonnormal distributions. This expansion is crucial to estimating the robustness of statistics under a variety of True Negative (TN) conditions. When doing so, researchers should employ the checking method we mentioned earlier to ensure that the design excludes any relationship specific to the statistic being tested.

Step 2: Suppose there is a commonly used cut-off or an acceptable level of a statistic with a continuous result. FPm (5) and FNm (6) should be measured in various conditions, such as those conditions commonly occurring in practice.

We have already described the difficulty of practicing FN and FP directly in statistics used in psychology. Yet, we also admit that FNm is necessary but not sufficient to estimate FN. Analogically speaking, using FNm to replace FN and using FPm to replace FP would be like trying to measure whether an unknown number X is bigger than 1 to solve the question of whether $X > 2$. If $X \leq 1$, then X is definitely less than 2. However, if $X > 1$, it does not necessarily mean X is greater than 2.

The confusion matrix design also works in this way. Suppose a statistic can report a result above the cut-off or an acceptable level of a relationship between variables measured by this statistic when there is a null distribution of the statistics in this condition. In this case, it is also highly likely that the statistic will report a result above this cut-off when the population parameter is lower than the acceptable level. As a result, the statistics in this condition are not reliable. To estimate the possibility of this situation, we conducted another simulation study that used internal consistency levels of 0.3 and 0.5 as the true parameters of the population. The result is in Table 5. According to our findings, the Omega is also boosted in the conditions tested as questionable by FPm. Therefore, FPm scores above 5% are reliable enough to ascertain when a statistic should be considered questionable. Some researchers might argue that this part of the simulation may also be included in our proposed confusion matrix design. Yet, for some statistics, it is not easy to find a present but not acceptable level of the statistic.

Furthermore, our research identified two key relationships between True Negative (TN) and FPm. If a statistic shows poor performance in the TN condition, it is likely to also fare poorly in the FPm condition. This observation aligns with the rationale we discussed earlier. Additionally, we found that a positive bias in TN is correlated with an increased likelihood of simulation study results meeting the acceptable cut-off. Using the original study as an empirical example of True Positive (TP), we can reasonably infer that all four indices demonstrate robust performance in FNm. Thus, for statistics without a pre-established cut-off, we recommend using TN and TP as predictive references. A large absolute value in percentage bias and RMSE suggests that the statistical output is likely derived from population samples.

Several research scenarios

We have demonstrated a comprehensive example of applying this enhanced confusion matrix design in evaluating internal consistency indices. To further clarify, we propose that this design is versatile and can be applied to a broader range of tasks. Before delving into a general discussion, we will present three concise examples illustrating how the confusion matrix design can be implemented in other published simulation studies. In the first two studies, only TN conditions can be applied, as these studies do not have a common cut-off for their respective statistics (i.e., correlation coefficients and mediation correlation coefficients). However, for the third study, we will apply the full confusion matrix design, as it involves a cut-off for Root Mean Square Error of Approximation (RMSEA) in Structural Equation Modeling (SEM).

Ventura-León et al. (2023) executed a Monte Carlo simulation study focusing on correlation coefficients commonly used in psychology research. They examined various population correlation conditions, such as 0.12, 0.20, 0.31, and 0.50, under nonnormal distributions and distributions with outliers. Their

TABLE 5 Estimation of false positive method with unacceptable internal consistency level.

QL	SS	Condition	IL	Alpha	Omega	GLB	GLBa
6	20	TE	0.3	0.75%	**56.45%**	**35.83%**	**34.83%**
6	25	TE	0.3	0.20%	**39.08%**	**23.99%**	**25.93%**
6	30	TE	0.3	0.12%	**27.14%**	**16.02%**	**19.77%**
6	35	TE	0.3	0.02%	**18.08%**	**10.25%**	**14.81%**
6	40	TE	0.3	0.01%	**12.68%**	**6.28%**	**10.77%**
6	45	TE	0.3	0.02%	**8.89%**	4.12%	**8.35%**
6	50	TE	0.3	0.00%	**6.63%**	2.60%	**6.68%**
6	250	TE	0.3	0.00%	0.00%	0.00%	0.00%
6	500	TE	0.3	0.00%	0.00%	0.00%	0.00%
6	1,000	TE	0.3	0.00%	0.00%	0.00%	0.00%
12	20	CG	0.5	6.82%	**60.38%**	**99.63%**	**92.11%**
12	25	CG	0.5	3.65%	**40.39%**	**98.48%**	**88.77%**
12	30	CG	0.5	2.07%	**28.16%**	**96.93%**	**85.42%**
12	35	CG	0.5	1.21%	**21.08%**	**95.50%**	**82.38%**
12	40	CG	0.5	1.07%	**16.04%**	**93.19%**	**79.36%**
12	45	CG	0.5	0.60%	**12.09%**	**90.52%**	**76.02%**
12	50	CG	0.5	0.47%	**9.49%**	**87.67%**	**73.44%**
12	250	CG	0.5	0.00%	0.00%	2.18%	**18.57%**
12	500	CG	0.5	0.00%	0.00%	0.00%	7.78%
12	1,000	CG	0.5	0.00%	0.00%	0.00%	3.43%

QL is the length of the questionnaire or the item number in a questionnaire; SS is the sample size. TE is tau-equivalent model. CG is Congeneric model. IL is the population internal consistency parameter of Alpha. Percentage values are failures that suggest that a statistic report that internal consistency is above the cut-off when in fact, there is an internal consistency parameter that is considerably away from this cut-off.

findings indicated that the Winzorized Pearson correlation coefficient (Wilcox, 2011) performed the best within the simulated conditions they included. Based on the design of our study, we suggest that Ventura-León et al. (2023) should also consider including conditions with a null distribution of the statistics, specifically where the population correlation is zero that can be used as TN. The absence of TN in their study leaves a gap in empirical evidence regarding the performance of correlation coefficients under this condition. This omission poses a risk, as certain correlation coefficients may exhibit poor performance at the zero point, like the Eta square effect size (Okada, 2013) and the Omega statistics in our simulation.

Sim et al. (2022) conducted a Monte Carlo simulation study to estimate the necessary sample size for detecting mediation effects in various models. Their study included partial and full mediation conditions, providing the minimum sample size required to detect these effects. However, their design overlooked the inclusion of null distribution of mediation effects conditions, which are crucial for assessing the sample size needed to maintain a reasonable Type-I error level. This omission can bring significant problems. For instance, suppose a sample size requirement of 200 is found under some null distribution conditions to ensure the correct result is found in most replications. Then, the conclusions of Sim et al. (2022) might be called into question. They concluded that a sample size of 90 is sufficient to detect a mediation effect when the factor loading is 0.7 with a large effect size. Yet, this sample size level may not avoid the detection of a mediation effect in a population where no such effect exists. Including conditions with no mediation effect, as TN proposed in our study, is essential to test and validate the sample size requirements thoroughly.

In the case of the studies by Ventura-León et al. (2023) and Sim et al. (2022), the simulation conditions of FPm and FNm are not applicable, as these studies lack defined criteria for determining satisfactory levels of mediation effect or correlation coefficients. Next, we will examine another study by Gao et al. (2020), which focuses on the RMSEA in SEM. Our discussion will first detail the design of Gao, Shi, and Maydeu-Olivares's study, followed by its shortcomings. We will then explore how the methodology of our study can be applied to theirs to address these limitations.

Gao et al. (2020) used a Monte Carlo simulation study to examine the robustness of several RMSEA measurements. Their studies have included several robust RMSEA measurement methods and conditions with normal and nonnormal distributions. They found that RMSEA with mean and variance corrections is the most robust as it performs best across all conditions.

From our perspective, the study conducted by Gao et al. (2020) has shortcomings. One significant limitation is their failure to test the statistics under a null distribution condition, such as a simulated distribution in which items bear no relationship to the model. This omission means that they have not provided empirical evidence about the performance of these statistics in such a null condition. Therefore, it is essential to include TN conditions in their analysis. Additionally, they should test whether any RMSEA correction methods can yield results considered a good fit under null distribution conditions. This FPm design could be assessed using a cut-off of 0.08, as Fabrigar (1999) suggested, across various conditions. If certain conditions reveal a good fit using an RMSEA correction method, then the performance of these statistics under these specific conditions becomes questionable. A similar approach could be applied to assess FNm.

General discussion

This study introduces a novel simulation design based on a confusion matrix framework. As we propose, this innovative design is particularly suited for use in simulation studies that focus on comparing the performance of statistical methods under various conditions. To demonstrate its applicability, we have presented three potential scenarios and a detailed example illustrating the implementation of this design.

It is somewhat surprising that researchers might overlook the fact that studies like the original one can only yield empirical evidence when the attribute under investigation reaches an acceptable level. Consider a hypothetical scenario where all populations in psychological research exhibit an acceptable level of a particular statistical parameter. In such a case, regardless of whether the original study violated any assumptions, there would be no necessity to develop statistics to verify the existence of an effect. Furthermore, it's important to reiterate that APA guidelines advise researchers to include a null effect in any simulation of effect, specifically the absence of a mean difference between two groups (Fan, 2012). However, the rationale provided by the APA primarily aims to prevent Type-I errors, potentially leading researchers to mistakenly believe that the null distribution of statistics is only relevant for inferential statistics closely related to NHST. Our research findings suggest otherwise; different statistics may perform variably under different conditions. Identifying the most suitable statistic for these conditions requires including these conditions with an evaluation of the commonly used criteria.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YC: Data curation, Writing – original draft. PP-D: Writing – review & editing. KP: Writing – review & editing. JL: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1298534/full#supplementary-material

References

Beaujean, A. A. (2018). Simulating data for clinical research: a tutorial. *J. Psychoeduc. Assess.* 36, 7–20. doi: 10.1177/0734282917690302

Bernaards, C., and Jennrich, R. (2015). Package "GPArotation." Available at: http://ftp.daum.net/CRAN/web/packages/GPArotation/GPArotation.pdf

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., and Hilgard, J. (2019). Correcting for bias in psychology: a comparison of meta-analytic methods. *Adv. Methods Pract. Psychol. Sci.* 2, 115–144. doi: 10.1177/25152459198471

Cortina, J. M., Sheng, Z., Keener, S. K., Keeler, K. R., Grubb, L. K., Schmitt, N., et al. (2020). From alpha to omega and beyond! A look at the past, present, and (possible) future of psychometric soundness in the journal of applied psychology. *J. Appl. Psychol.* 105, 1351–1381. doi: 10.1037/apl0000815

Croasmun, J. T., and Ostrom, L. (2011). Using likert-type scales in the social sciences. *J. Adult Educ.* 40, 19–22.

Derrick, B., Toher, D., and White, P. (2016). Why Welch's test is type I error robust. The quantitative methods. *Psychology* 12, 30–38. doi: 10.20982/tqmp.12.1.p030

Fabrigar, L. R. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods* 4, 272–299. doi: 10.1037/1082-989X.4.3.272

Fan, X. (2012). "Designing simulation studies" in *APA handbook of research methods in psychology*. eds. H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf and K. J. Sher (Washington, DC: Data analysis and research publication, American Psychological Association). *Vol. 3*. 427–444.

Fan, X., and Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivar. Behav. Res.* 42, 509–529. doi: 10.1080/00273170701382864

Fernández-Castilla, B., Declercq, L., Jamshidi, L., Beretvas, S. N., Onghena, P., and Van den Noortgate, W. (2021). Detecting selection bias in meta-analyses with multiple outcomes: a simulation study. *J. Exp. Educ.* 89, 125–144. doi: 10.1080/00220973.2019.1582470

Fisk, C. L., Harring, J. R., Shen, Z., Leite, W., Suen, K. Y., and Marcoulides, K. M. (2023). Using simulated annealing to investigate sensitivity of SEM to external model misspecification. *Educ Psychol Meas*. 83, 73–92.

Gao, C., Shi, D., and Maydeu-Olivares, A. (2020). Estimating the maximum likelihood root mean square error of approximation (RMSEA) with nonnormal data: a Monte-Carlo study. *Struct. Equ. Model. Multidiscip. J.* 27, 192–201. doi: 10.1080/10705511.2019.1637741

Gelfand, A., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* 85, 398–409. doi: 10.1080/01621459.1990.10476213

Greene, A. L., Eaton, N. R., Li, K., Forbes, M. K., Krueger, R. F., Markon, K. E., et al. (2019). Are fit indices used to test psychopathology structure biased? A simulation study. *J. Abnorm. Psychol.* 128, 740–764. doi: 10.1037/abn0000434

Hayes, A. F., and Counts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Commun. Methods Meas.* 14, 1–24. doi: 10.1080/19312458.2020.1718629

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Comp. Statis. Data Anal.* 40, 685–711. doi: 10.1016/S0167-9473(02)00072-5

Heggestad, E. D., Rogelberg, S., Goh, A., and Oswald, F. L. (2015). Considering the effects of nonresponse on correlations between surveyed variables: a simulation study toprovide context to evaluate survey results. *J. Pers. Psychol.* 14, 91–103. doi: 10.1027/1866-5888/a000129

Hunter, M.A., and May, R. B. (2003). Statistical testing and null distributions: What to do when samples are not random. *Can. J. Exp. Psychol.* 57, 176–188.

Marom, N. D., Rokach, L., and Shmilovici, A. (2010). Using the confusion matrix for improving ensemble classifiers, 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, 000555–000559.

McDonald, R. (1999). *Test theory: A unified treatment*. Hillsdale, NJ: Lawrence Erlbaum.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychol. Methods* 23, 412–433. doi: 10.1037/met0000144

Moltner, A., and Revelle, W. (2015). Find the greatest lower bound to reliability. Available at: http://personality-project.org/r/psych/help/glb.algebraic.html

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* 50, 885–900. doi: 10.13031/2013.23153

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika* 40, 129–147. doi: 10.2333/bhmk.40.129

Okada, K. (2017). Negative estimate of variance-accounted-for effect size: how often it is obtained, and what happens if it is treated as zero. *Behav. Res. Methods* 49, 979–987. doi: 10.3758/s13428-016-0760-y

Patz, R., and Junker, B. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *J. Educ. Behav. Stat.* 24, 146–178. doi: 10.2307/1165199

Patz, R., and Junker, B. (1999b). Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *J. Educ. Behav. Stat.* 24, 342–366. doi: 10.2307/1165367

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Racine, J. (2012). Rstudio: a plataform-independet IDE for R and sweave. *J. Appl. Econ.* 27, 167–172. doi: 10.1002/jae.1278

Revelle, W. (2015). *Psych: Procedures for personality and psychological research*, Northwestern University, Evanston, Illinois, USA.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P., and Mononen, J. (2018). Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. *Behav. Process.* 148, 56–62. doi: 10.1016/j.beproc.2018.01.004

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 107–120. doi: 10.1007/s11336-008-9101-0

Sim, M., Kim, S. Y., and Suh, Y. (2022). Sample size requirements for simple and complex mediation models. *Educ. Psychol. Meas.* 82, 76–106. doi: 10.1177/00131644211003261

Spurrier, J. D. (2003). On the null distribution of the Kruskal–Wallis statistic. *Nonparamet. Stat.* 15, 685–691. doi: 10.1080/10485250310001634719

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fittest statistic in IRT models. *J. Educ. Meas.* 37, 58–75. doi: 10.1111/j.1745-3984.2000.tb01076.x

Streiner, D. L., and Cairney, J. (2007). What's under the ROC? An introduction to receiver operating characteristics curves. *Can. J. Psychiatry* 52, 121–128. doi: 10.1177/070674370705200210

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Res. Sci. Educ.* 48, 1273–1296. doi: 10.1007/s11165-016-9602-2

Tang, W., Cui, Y., and Babenko, O. (2014). Internal consistency: Do we really know what it is and how to assess it. *J. Psychol. Behav. Sci.* 2, 205–220.

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

Ventura-León, J., Peña-Calero, B. N., and Burga-León, A. (2023). The effect of normality and outliers on bivariate correlation coefficients in psychology: a Monte Carlo simulation. *J. Gen. Psychol.* 150, 405–422. doi: 10.1080/00221309.2022.2094310

Wang, W., and Lu, Y. (2018). "Analysis of the mean absolute error (MAE) and the root mean square error (RMSE) in assessing rounding model" in *IOP Conference Series: Materials Science and Engineering* (Kuala Lumpur, Malaysia: IOP Publishing)

Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: from alpha to omega. *Clin. Neuropsychol.* 31, 1113–1126. doi: 10.1080/13854046.2017.1317364

Wilcox, R. (2011). *Modern statistics for the social and behavioral sciences: a practical introduction* (*2nd*). Boca Raton: CRC Press.

Appendix

The R code that generated all the data and simulation results in this study is available in a separate file that is attached to the current submission to the journal Frontiers in Psychology. It is also available through the URL: https://liqas.org/code-under-review/. Researchers are encouraged to simulate and replicate the results for future research. This study was not preregistered.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 05 January 2024
DOI 10.3389/fpsyg.2023.1268283

OPEN ACCESS

EDITED BY
Georgios Sideridis,
Harvard Medical School, United States

REVIEWED BY
Cristian Ramos-Vera,
Cesar Vallejo University, Peru
Heinz Leitgöb,
Leipzig University, Germany

*CORRESPONDENCE
Daniel Zimprich
✉ daniel.zimprich@uni-ulm.de

RECEIVED 27 July 2023
ACCEPTED 11 December 2023
PUBLISHED 05 January 2024

CITATION
Zimprich D, Pociūnaitė J and Wolf T (2024) A multilevel factor analysis of the short form of the Centrality of Event Scale.
Front. Psychol. 14:1268283.
doi: 10.3389/fpsyg.2023.1268283

A multilevel factor analysis of the short form of the Centrality of Event Scale

Daniel Zimprich*, Justina Pociūnaitė and Tabea Wolf

Department of Developmental Psychology, Institute of Psychology and Education, Ulm University, Ulm, Germany

Introduction: The Centrality of Event Scale (CES) has frequently been used to measure the degree to which positive and negative life events are perceived central to a person's identity and life story; and previous research suggests that individuals rate their most positive memory as more central compared to their most negative one. When comparing the centrality of two (or more) memories within individuals, one needs to ensure that the CES (or its short form) is equally valid for different types of events (i.e., positive and negative) as well as on different levels of analyses (i.e., on the between-person and the within-person level), pointing to the issue of measurement invariance.

Methods: Three-hundred sixty-five adults (18–89 years of age) reported up to ten positive and up to ten negative autobiographical memories. For each memory reported, participants completed the seven-item short form of the CES, which measures three different components of centrality: Events can form a central component of identity (two items), a turning point in the life story (three items), and a reference point for everyday inferences (two items).

Results: Based on exploratory and confirmatory factor analyses, we found a two-factor structure (Self-Perception and Life-Course) to fit the data best at both levels of analyses and for both positive and negative events. Strict measurement invariance could be applied for positive and negative events at between-person level and at within-person level. The two factors, which measure the impact of an event on either a person's self-perception or their (future) life course, were rated higher for positive compared to negative memories. This difference, however, was stronger for the self-perception factor.

Discussion: The present study provides a first examination of the factorial structure of the CES short form on two levels (within and between persons) as well as for two types of life events (positive and negative). Whereas, a unidimensional scale might be sufficient to measure the centrality of stressful or traumatic life events, a more fine-graded measure seems better suited to understand the different roles of positive and negative life events for a person's identity and life story.

KEYWORDS

Centrality of Event Scale, multilevel factor analysis, positive autobiographical memories, negative autobiographical memories, within persons, between persons

1 Introduction

The Centrality of Event Scale (CES), originally developed to measure whether and, if so, to which degree stressful and traumatic life events have become central to an individual's identity and life story (Berntsen and Rubin, 2006), has extensively been used in research on post-traumatic stress disorder (e.g., Schuettler and Boals, 2011; Groleau et al., 2013), psychopathology (e.g., Pinto-Gouveia and Matos, 2011), and depression (e.g., Newby and Moulds, 2011). Results of these studies have shown that the centrality of a stressful and traumatic event is correlated with the severity of symptoms in post-traumatic stress disorder (Brown et al., 2010), prolonged grief disorder and depression (Boelen, 2012), current feelings of shame (Pinto-Gouveia and Matos, 2011), and poor physical health outcomes (Boals, 2010).

In the present study, our goal was to shed some light on the measurement properties of (the brief version of) the CES (see below). More specifically, we aimed to clarify whether event centrality can be measured comparably for positive and negative events as well as at two levels of data—between persons and within persons.

1.1 Factorial structure of the CES

Although the CES has frequently been used, its factorial structure has been investigated only a few times. Originally, the CES was developed to assess three different possible functions that (traumatic) life events may have (Berntsen and Rubin, 2006). The first function entails how a (traumatic) life event has become a reference point, which, from a functional memory perspective, serves as guidance for future behavior, or for learning from one's past experiences (Pillemer, 2009; Rasmussen and Berntsen, 2009). An exemplar item from the CES capturing this function is "This event has become a reference point for the way I understand new experiences." A second function captures how a (traumatic) life event is seen as a turning point in one's life. From a life narrative perspective, the traumatic event thus functions as closing one chapter and beginning another (e.g., Habermas, 2019). An item from the CES that reflects this function is "If this event had not happened to me, I would be a different person today." Finally, the third function addresses how an event has become a part of one's personal identity, such that the event is seen as a symbol or theme in one's life. The CES taps this phenomenon through items such as "I automatically see connections and similarities between this event and experiences in my present life."

In line with these considerations, an exploratory factor analysis of the CES in an undergraduate sample returned three factors with eigenvalues larger than one (Berntsen and Rubin, 2006). However, because there was a drop in the size of eigenvalues from the first compared to the other two eigenvalues, the authors proposed the 20 items of the CES to be unidimensional, that is, to measure *one* underlying latent variable (or factor) of event centrality. Unfortunately, indexes of model fit, factor loadings, or measures of explained variance were not reported, such that the adequacy of a one-factor model compared to a three-factor model cannot be fully evaluated.

By contrast, in a sample of 195 Brazilian undergraduate students, Gauer et al. (2013) found the 20-item CES to be composed of three orthogonal factors, which they found via exploratory factor analysis followed by varimax rotation. Similar to Berntsen and Rubin (2006), there was a drop in eigenvalues from the first eigenvalue on, but the authors nevertheless opted for a three-factor solution. The interpretation of the three factors was in line with the functions proposed theoretically. Specifically, the first factor, on which 10 items showed loadings >0.45, was interpreted as the extent to which the memory of an event has become a reference point for everyday life. The authors interpreted the second factor, on which seven items had loadings >0.45, as the degree the memory of an event has turned into a central component of a person's identity. Finally, the third factor (three loadings >0.45) measured the amount of which an event reflected a turning point in a person's life story.[1] Note that the factors were chosen to be mutually uncorrelated (i.e., orthogonal), which, in turn, implies that items can show loadings on all three factors. Because the authors decided to only report factor loadings >0.45, the interpretation of the factors is not completely transparent, since it remains unknown whether items significantly loaded on more than one factor and, if so, how strong these cross-loadings were. Moreover, indexes of model fit were not given in the article.

In a sample of 872 Italian adolescents, Ionio et al. (2018) also found a three-factor solution using confirmatory factor analysis, which mapped the factors proposed theoretically by Berntsen and Rubin (2006). The first factor, on which eight items were designated to load (loadings ranging from 0.60 to 0.78), assessed the extent to which an event had become a reference point for expectations and the attribution of meaning to other personal life events. The second factor, composed of seven items with loadings ranging from 0.65 to 0.78, measured the perception of an event as central to one's personal identity. Finally, the third factor (five items with loadings ranging from 0.73 to 0.83) reflected whether an event was perceived as a turning point in one's life story. In addition, the authors tested for measurement invariance across gender and found that factor loadings and intercepts were equal for females and males, implying strong measurement invariance (Meredith and Horn, 2001). Some relevant information is missing in the article, however. For example, the correlations among factors were not reported. In addition, after having established strong measurement invariances across gender, differences in factor parameters (variances, covariances, and means) have, apparently, not been analyzed.

In a study of 1,079 Portuguese adolescents, Vagos et al. (2018) found a three-factor solution as well, which was based on item content and achieved the best model fit. The first factor ("reference point", on which 7 items loaded) was similar to that of Ionio et al. (2018). Likewise, the second factor ("turning point", five items) and the third factor ("personal identity", six items) showed substantial overlap with the Ionio et al. (2018) solution. Notably, however, the authors excluded Items 2 and 11, such that the analyses were

1 The amount of explained variance by the respective factors as given in Table 1 in Gauer et al. (2013) was, obviously, calculated before the varimax rotation. Notwithstanding, from these numbers the total sum of variance explained by the three-factor solution can be calculated, which gives 62%.

based on 18 items. Strong measurement invariance across female and male subsamples was established and a subsequent comparison of factor means showed that females had lower means on factors 2 and 3. The three factors were strongly correlated, ranging from 0.74 between "reference point" and "personal identity" to 0.85 between "reference point" and "turning point".

In a sample of 263 adults who had experienced at least one traumatic event, Wamser-Nanney (2019) reported that a CES three-factor solution fit the data adequately. However, the three factors were very strongly correlated (r = 0.92–0.96), wherefore the author conducted further analyses with a one-factor model—albeit the one-factor solution only showed a marginal fit for the data and represented a significant decrease in fit compared to the three-factor model.

In a recent article, Bruce and Handal (2023) examined the CES factor structure in a sample of MTurk participants recruited online for a survey-based study on self-reported experiences post-trauma and a sample of students exposed to trauma. For the data analysis, from both studies those participants who described their trauma as either bereavement (N = 221) or sexual assault (N = 97) were selected, resulting in a sample size of 318 persons. In both groups, a two-factor solution emerged from an exploratory factor analysis using varimax rotation. Notwithstanding, in both groups a one-factor solutions was also evaluated, which accounted for 54 and 61% of variance, respectively. Indexes of model fit were not reported.

To summarize previous research on the factorial structure of the CES, it appears as if three factors may be more appropriate to describe the associations among the 20 items—at least in samples of younger adults and predominantly regarding traumatic, stressful or (the most) negative life events. Moreover, the three-factor solutions, with a grain of salt, map to the theoretical structure suggested by Berntsen and Rubin (2006). At the same time, factors are typically strongly correlated, which is why some authors opted for a one-factor solution (e.g., Wamser-Nanney, 2019; Bruce and Handal, 2023). Note, however, that the strong correlations among factors (which imply strong inter-item correlations) may result from the fact that mostly traumatic and most negative events were evaluated by participants—one would expect relatively strong endorsement of all CES items in this case. What complicates a thorough evaluation of previous studies on the factorial structure of the CES is that different analysis approaches have been used (e.g., orthogonal vs. oblique rotation), results stem from samples differing in the severity of the events evaluated using the CES, different language versions of the CES have been employed, and, finally, relevant information is missing in publications.

1.2 The CES short form

Berntsen and Rubin (2006) also suggested a brief version of the CES, composed of those seven items that were most strongly correlated with the total score of the original scale. This brief version has also frequently been used in research on the centrality of life events (e.g., Boals, 2010; Rubin et al., 2014). Only a few studies have examined the factorial structure of this short form and, again, these studies focused exclusively on traumatic, respectively negative life events of young adults. Most of them favor a single factor structure as proposed by Berntsen and Rubin (2006).

For instance, in the aforementioned study, Vagos et al. (2018) not only investigated the factorial structure of the full version, but also of the short form of the CES. The authors specified three measurement models, the unidimensional model suggested by Berntsen and Rubin (2006), a unidimensional model based on the seven strongest items correlations suggested by Gauer et al. (2013), and, finally, a three-factor model representing the theoretically postulated components of the CES (i.e., reference point, turning point, and personal identity). The authors favored the unidimensional solution suggested by Gauer et al. (2013), although the three-factor model showed a better fit in terms of RMSEA, CFI, and SRMR. One has to keep in mind, though, that the short form is comprised of seven items only, implying that either one or two factors can be extracted in a meaningful way (based on the requirement of a minimum of three indicators per factor).

Galán et al. (2017) also tested the factorial structure of both the full and short version of the CES in a sample of undergraduates from Spain. Based on two confirmatory factor analyses, their findings support a single factor structure for both CES versions. It is unclear, however, whether other CFA models with more than one factor were tested, because the authors reported results for the single factor solutions only. The same holds for a study conducted by Azadfar et al. (2022). These authors tested the unidimensional structure of the CES short form (and only the single factor structure) in a sample of Iranian university students with a history of at least one romantic breakup, on which the CES measure was based on. Measurement invariance analyses showed that the single factor structure of the CES short form was invariant across gender.

Vermeulen et al. (2020) based their analyses on a sample of 311 Dutch-speaking psychology students (mostly female). Their data favor a single factor solution based on a factor analysis for ordered-categorical data. However, the authors found the best fit for a model that is comprised of six items only (excluding item 7: "This event was a turning point in my life").

With respect to the CES short form, results of previous studies appear much more unequivocal. In general, a one-factor solution seems to capture the associations among the seven items adequately. However, as for the full CES, relevant information that would help evaluate findings more carefully is lacking in almost all studies.

1.3 Centrality of positive and negative events

More recently, the CES (most frequently in its brief version) has also been applied to assess the event centrality of non-traumatic autobiographical events, for example, positive vs. negative life events. Based on the so-called "positivity bias" in autobiographical memory (Walker et al., 2003), individuals are expected to focus on positive information about their personal past more strongly than on negative information. Similarly, the "fading affect bias" (Walker et al., 1997) suggests that the affect intensity of negative events decreases more quickly across time than the affect intensity of positive events (see Hoehne, 2023). The assumption thus

is that individuals tend to assign stronger centrality ratings to emotionally positive events compared to emotionally negative events (Pociūnaitė and Zimprich, 2023).

In line with this assumption, Berntsen et al. (2011) found that in older persons the centrality ratings differed in dependence on whether the life event was positive or negative, with the former having a significantly higher event centrality. Similar findings were reported by Zaragoza Scherman et al. (2015). Their study included middle-aged and older adults from Mexico, Greenland, China, and Denmark. Participants completed event centrality scales for their most positive and most negative life events. Across cultures, participants rated positive events as more central than negative events. The same authors conducted a similar study to compare centrality ratings for highly positive and highly negative memories in a sample including young and middle-aged adults, again from Mexico, Greenland, China, and Denmark (Zaragoza Scherman et al., 2020). Both age groups rated their positive memories as more central compared to their negative memories. However, the relative difference between those ratings was smaller in the young adults group (younger adults reported a lower centrality of positive memories than middle-aged adults did). This aligns with studies focusing on samples of younger adults that found no differences in the event centrality ratings between positive and negative events (see Rasmussen and Berntsen, 2009; Boals, 2010, but see Rasmussen and Berntsen, 2013).

Note that one precondition to compare the centrality of emotionally positive vs. emotionally negative autobiographical events is that the CES (or its short form, which was mainly used in previous studies) is equally valid for both types of events. If this precondition does not hold, observed score differences (i.e., CES means of positive vs. negative events) will not accurately reflect true differences in the quantity being measured (i.e., centrality). Psychometricians have developed theory and methods for assessing whether scores are equivalent in meaning and metric across individuals and/or within individuals (e.g., judging the centrality of positive vs. negative events), a condition referred to as measurement invariance (Meredith, 1993; Meredith and Horn, 2001). What we refer to here is not measurement invariance between (groups of) persons—something that has already been examined by Vagos et al. (2018), for example, with respect to males and females. Our concern here is measurement invariance within persons, that is, whether centrality is measured in a comparable manner for positive and negative events when individuals rate centrality for both event types.

1.4 The present study: a multilevel perspective on event centrality

In the present study, we approach the measurement of event centrality from two different, but related perspectives, a within-person and a between-person perspective. Moreover, these two perspectives will be adopted for both positive and negative events (cf. Pociūnaitė et al., 2022).

The measurement of event centrality can help answering two conceptually different questions. The first question touches upon the measurement of *differences between persons* in the sense of, for example, examining whether persons with post-traumatic stress disorder symptoms judge the centrality of a stressful event higher than persons with no post-traumatic stress disorder symptomatology. This type of investigation, which can be described as examining between-person or *inter*individual differences in event centrality, is the predominant way the CES has been used in previous studies (e.g., Ionio et al., 2018; Bruce and Handal, 2023).

There is a second perspective on event centrality. If participants are asked, for example, to judge the centrality of events forming their emotionally most positive vs. their emotionally most negative autobiographical memories, the measurement of event centrality can also refer to *within-person* or *intra*individual differences, that is, differences among events. For example, the event centrality of an emotionally negative event might be higher within individuals than that of an emotionally positive event (e.g., Zaragoza Scherman et al., 2015). This within-person perspective comes into play as soon as participants are asked to rate the event centrality of more than one event from their past.

These two types of measuring event centrality—one within-person, the other between-person—can be systematically compared with respect to their measurement qualities by imposing different degrees of measurement invariance (see below). Even more options to examine measurement invariance come into play when the within- and between-person perspectives are transferred to event centrality measurements of positive vs. negative events.

More specifically, in the present study we address the following research questions: (1) Is the measurement of event centrality (as measured by the brief CES) comparable for positive and negative events? (2) Is the measurement of event centrality comparable within and between persons? (3) Combining questions (1) and (2), is the measurement of event centrality comparable both for positive and negative events *and* within and between persons?

2 Methods

2.1 Sample

The sample of the present study comprised 365 adults aged between 18 and 89 years ($M = 49.58$, $SD = 17.05$).[2] The majority of the sample was female (67.1%). Participants were mostly married (58.6%) or single (28.8%). Almost half of the sample had graduated from university (45.2%). Sixty-two participants were university students (17%). Most of them belonged to the group of young adults ($n = 60$). The majority of the sample reported to be employed, but occupational status differed considerably with age. Overall, subjective health was rated as good ($M = 2.23$, $SD = 0.88$) on a scale ranging from excellent (1) to poor (5).

Participants were recruited through promotional flyers, e-mail, and word of mouth. To participate in the study, individuals had to be at least 18 old and have a good working knowledge of the German language. After finishing the study, they could take part in a lottery to win a gift voucher (worth 15 Euros). For students, there was an option to get course credit (instead of lottery).

2 Part of the data have been used in a previous study with a different focus (see Pociūnaitė et al., 2022).

2.2 Procedure and measures

Data were collected online using the www.soscisurvey.de platform (Leiner, 2019). After having given their informed consent, participants provided demographic information (e.g., age, gender, marital status, education) and rated their subjective health. Next, participants were asked to recall up to ten positive memories. They were instructed to briefly describe the (first) memory that came to their mind. Participants were told that memories did not have to be extraordinary, but should refer to a specific and distinct event from their personal past. For each memory, a separate page was provided where participants were asked to enter a brief description of the event and proceed to the next memory once they were finished. In the next step, participants were asked to recall up to ten negative memories. The instruction and the procedure were identical to the one for positive memories. If participants did not find 10 positive and/or 10 negative memories to report, they could proceed to the next page. Order of the procedure was the same for all participants.

After having described positive and negative memories, participants completed a personality questionnaire. Subsequently, participants were presented with their description of positive and negative memories and were asked to answer several questions concerning the events described (see below). Memories were presented in the order in which they had been recalled (again, starting with positive and then negative memories).

Centrality of event. Participants rated the event centrality for each reported memory. We used the seven-item short version of the CES, which—as suggested by Berntsen and Rubin (2006)—consists of Items 3, 6, 10, 12, 16, 17, and 18 of the original CES. Responses were made on a 5-point Likert-scale ranging from totally disagree (1) to totally agree (5). German item wordings were based on the translation of two independent researchers and are very similar to those of the recently published German version of the full CES (Conen et al., 2022).

2.3 Modeling approach

The data in the present study represent a typical multilevel situation, where measurements (centrality of event of different positive and negative autobiographical memories) are nested within persons (Hox, 1995). Consider a multivariate situation of multilevel data, in which there are $i = 1, \ldots, N$ individuals (Level 2) and within each individual, there are p variables (i.e., the seven CES items) measured with respect to $j = 1, \ldots, m_i$ autobiographical memories (Level 1).[3] Let $\mathbf{y}_{ij}$ denote the $p \times 1$ vector of CES items measured in individual i with respect to autobiographical memory j. Suppose that this vector of measured variables is composed as

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \mathbf{v}_i + \mathbf{w}_{ij},$$

where $\boldsymbol{\mu}$ is a $p \times 1$ vector of overall (or sample) means of the CES items, $\mathbf{v}_i$ is a $p \times 1$ vector of deviations of the individual-specific means of the CES items from the overall means (i.e., $\mathbf{v}_i = \bar{\mathbf{y}}_i - \boldsymbol{\mu}$, where $\bar{\mathbf{y}}_i$ denotes the vector of individual-specific means of the CES items) and $\mathbf{w}_{ij}$ is a $p \times 1$ vector of memory-specific deviations from the individual-specific mean deviations (i.e., $\mathbf{w}_{ij} = \mathbf{y}_{ij} - \bar{\mathbf{y}}_i$). The vectors $\mathbf{v}_i$ and $\mathbf{w}_{ij}$ are independent with expectations $\mathcal{E}(\mathbf{v}_i) = \mathcal{E}(\mathbf{w}_{ij}) = \mathbf{0}$ and covariance matrices $\mathcal{C}(\mathbf{v}_i) = \boldsymbol{\Sigma}_2$, the covariance matrix of *inter*individual (or between-person) differences, and $\mathcal{C}(\mathbf{w}_{ij}) = \boldsymbol{\Sigma}_1$, the covariance matrix of *intra*individual (or within-person) differences. Assume that the between-person or interindividual differences at Level 2 can be described by a factor analysis model (Longford and Muthén, 1992) such that

$$\mathbf{v}_i = \boldsymbol{\Lambda}_b \boldsymbol{\xi}_i + \mathbf{u}_i,$$

where $\boldsymbol{\Lambda}_b$ is a $p \times q$ matrix of factor loadings at Level 2 (or the between-person level), $\boldsymbol{\xi}_i$ is a $q \times 1$ vector of factor scores of individual i at Level 2, and $\mathbf{u}_i$ is a $p \times 1$ vector of residuals at Level 2. Factor scores are assumed to be normally distributed with zero means and covariance matrix $\boldsymbol{\Phi}$, that is, $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$. Similarly, residuals are normally distributed with zero means and covariance matrix $\boldsymbol{\Theta}_u$, that is, $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_u)$. Assuming that factor scores and residuals are independent, the between-person or Level 2 covariance matrix predicted by the factor analysis model is

$$\boldsymbol{\Sigma}_2 = \mathbf{v}_i \mathbf{v}_i' = \boldsymbol{\Lambda}_b \boldsymbol{\Phi} \boldsymbol{\Lambda}_b' + \boldsymbol{\Theta}_u.$$

Moreover, suppose that the within-person or intraindividual differences can also be described by a factor analysis model, that is,

$$\mathbf{w}_{ij} = \boldsymbol{\Lambda}_w \boldsymbol{\eta}_{ij} + \mathbf{e}_{ij},$$

where $\boldsymbol{\Lambda}_w$ is a $p \times r$ matrix of factor loadings at Level 1 (or the within-person level), $\boldsymbol{\eta}_{ij}$ is a $r \times 1$ vector of factor scores at Level 1, and $\mathbf{e}_{ij}$ is a $p \times 1$ vector of residuals at Level 1. Both factor scores and residuals at Level are assumed to be independent and normally distributed with zero means and covariance matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}_e$, respectively, that is, $\boldsymbol{\eta}_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ and $\mathbf{e}_{ij} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Theta}_e)$. The predicted Level 1 covariance matrix then is

$$\boldsymbol{\Sigma}_1 = \mathbf{w}_{ij} \mathbf{w}_{ij}' = \boldsymbol{\Lambda}_w \boldsymbol{\Psi} \boldsymbol{\Lambda}_w' + \boldsymbol{\Theta}_e.$$

The total covariance matrix of observed variables is thus equal to (cf. McDonald, 1993)

$$\boldsymbol{\Sigma}_{\text{total}} = \boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_1 = \underbrace{\boldsymbol{\Lambda}_b \boldsymbol{\Phi} \boldsymbol{\Lambda}_b' + \boldsymbol{\Theta}_u}_{\text{between-person}} + \underbrace{\boldsymbol{\Lambda}_w \boldsymbol{\Psi} \boldsymbol{\Lambda}_w' + \boldsymbol{\Theta}_e}_{\text{within-person}}. \quad (1)$$

The Level 2 or between-person part of Equation (1) is to be interpreted in line with conventional factor analysis, that is, the between-factors and the between-residuals refer to *inter*individual differences. The within part in Equation (1), however, differs from standard factor analysis in that it reflects the associations among *intra*individual differences (cf. Mehta and Neale, 2005). Here, factors capture shared within-person differences in judging the event centrality of different autobiographical memories.

3 The subscript i for the number m of AMs designates that individuals were allowed to differ in the number of AMs they reported.

TABLE 1 Four models of the CES estimated simultaneously.

	Positive (p) events	Negative (n) events
Level 1: within (w) persons	$\boldsymbol{\Sigma}_{1p} = \boldsymbol{\Lambda}_{wp}\boldsymbol{\Psi}_{\boldsymbol{p}}\boldsymbol{\Lambda}'_{wp} + \boldsymbol{\Theta}_{ep}$	$\boldsymbol{\Sigma}_{1n} = \boldsymbol{\Lambda}_{wn}\boldsymbol{\Psi}_{\boldsymbol{n}}\boldsymbol{\Lambda}'_{wn} + \boldsymbol{\Theta}_{en}$
Level 2: between (b) persons	$\boldsymbol{\Sigma}_{2p} = \boldsymbol{\Lambda}_{bp}\boldsymbol{\Phi}_{\boldsymbol{p}}\boldsymbol{\Lambda}'_{bp} + \boldsymbol{\Theta}_{up}$	$\boldsymbol{\Sigma}_{2n} = \boldsymbol{\Lambda}_{bn}\boldsymbol{\Phi}_{\boldsymbol{n}}\boldsymbol{\Lambda}'_{bn} + \boldsymbol{\Theta}_{un}$
Total covariance structure	$\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_{1p} + \boldsymbol{\Sigma}_{2p}$	$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}_{1n} + \boldsymbol{\Sigma}_{2n}$
Level 2: mean structure	$\boldsymbol{\mu}_p = \boldsymbol{\upsilon}_p + \boldsymbol{\Lambda}_{bp}\boldsymbol{\kappa}_p$	$\boldsymbol{\mu}_n = \boldsymbol{\upsilon}_n + \boldsymbol{\Lambda}_{bn}\boldsymbol{\kappa}_n$

2.4 Multilevel measurement invariance

Measurement invariance (MI) in general—and in a multilevel situation in particular—is a matter of degree (e.g., Zimprich et al., 2005, 2006, 2012; Zimprich and Martin, 2009; Wolf and Zimprich, 2015). More specifically, one may distinguish four forms of measurement invariance (cf. Meredith, 1993; Meredith and Horn, 2001). *Configural invariance* entails that the number of factors and the according salient and non-salient loadings are equal at both levels, i.e., within and between persons, which ensures that the dimensionality of the measured construct is equivalent. *Weak invariance* (or pattern invariance) requires that factor loading matrices be fully invariant within and between persons, i.e., $\boldsymbol{\Lambda}_w = \boldsymbol{\Lambda}_b$. On a conceptual level, weak invariance ensures that the same manifest variables (the seven CES items) relate to concepts (factors) in the same way. With weak MI holding, factor variances and covariances can be compared across levels, because the factors are scaled equally. *Strong invariance* (or metric invariance) requires that, in addition to factor loading matrices, latent intercepts of the manifest indicators be invariant. As such, because latent intercepts are only estimated at Level 2, it has no direct equivalent in a multilevel factor analysis. Finally, *strict invariance* adds the constraint of residual variances be invariant at both levels. Although, technically, it is possible to impose strict MI (more specifically, equal residual variances) in a multilevel factor analysis, one would typically not expect it to hold because on Level 2 residual covariances are typically much smaller because they represent "average" residual variances across Level 1 units.

As noted above, in the present study centrality of event was rated for up to 10 positive and up to 10 negative events. Comparing the measurement of event centrality across positive and negative events allows for more invariance analyses than by a typical multilevel factor analysis alone. If one combines the two-level data situation with the fact that centrality ratings were given for positive and negative events, a scheme of four (sub-)models emerges that can be examined with respect to their measurement properties. This scheme is shown in Table 1 with an obvious extension of notation using p for positive events and n for negative events.

Given this scheme, configural invariance can be investigated for (1) the measurement of event centrality positive and negative events, (2) for the measurement within and between persons, and (3) for positive and negative events *and* for both analysis levels. For weak invariance, the same three types of models can be examined, such that weak invariance can hold across levels, across positive and negative events, and both.

Regarding strong invariance—which cannot be tested across levels—we can investigate the equality of item intercepts across positive and negative events on Level 2 (between persons). This requires additional notation as shown in Table 1 under "mean structure." Here, $\boldsymbol{\mu}_p$ and $\boldsymbol{\mu}_n$ are the observed means for the CES items as rated for positive and negative events, $\boldsymbol{\upsilon}_p$ and $\boldsymbol{\upsilon}_n$ are the latent intercepts of the CES items for positive and negative events, and $\boldsymbol{\kappa}_p$ and $\boldsymbol{\kappa}_n$ are the factor means for positive and negative events. Strong invariance across positive and negative events then implies

$$\boldsymbol{\upsilon}_p = \boldsymbol{\upsilon}_n = \boldsymbol{\upsilon}.$$

Given that weak invariance also holds across positive and negative events, i.e., $\boldsymbol{\Lambda}_{bp} = \boldsymbol{\Lambda}_{bn} = \boldsymbol{\Lambda}_b$, we have that

$$\begin{aligned}\boldsymbol{\mu}_p - \boldsymbol{\mu}_n &= (\boldsymbol{\upsilon}_p + \boldsymbol{\Lambda}_{bp}\boldsymbol{\kappa}_p) - (\boldsymbol{\upsilon}_n + \boldsymbol{\Lambda}_{bn}\boldsymbol{\kappa}_n) \\ &= \boldsymbol{\Lambda}_b(\boldsymbol{\kappa}_p - \boldsymbol{\kappa}_n),\end{aligned}$$

which shows that, between persons, factor means can directly be compared across positive and negative events. One has to keep in mind that Level 2 is the only data level where factor mean comparisons appear meaningful, because on Level 1 a comparison of factor means of positive and negative events, if possible, would entail a comparison of 10 positive with 10 negative events or 45 comparisons in total.

2.5 Assessing model fit

Typically, the fit of an entire multilevel model is evaluated simultaneously—as it is done in ordinary confirmatory factor analysis, for example (e.g., Zimprich et al., 2005). In multilevel data, however, the sample size is usually much larger at Level 1 (within persons) compared to Level 2 (between persons). In our case, there were 5,081 events reported by 365 individuals. For this reason, the fit of the entire model is likely to be dominated by the (lack of) fit on Level 1 and may not be sensitive enough to model misspecifications at Level 2 (Yuan and Bentler, 2007; Ryu and West, 2009). To overcome this problem of standard fit indexes, two approaches have been developed to evaluate model fit in multilevel structural equation models. One approach utilizes partially saturated models to obtain the fit of the Level 1 and Level 2 models separately (Ryu and West, 2009). The other approach, in a first step, estimates the (asymptotic) covariance matrices of the manifest variables at Level 1 and Level 2, which are then used as input data in single-level structural equation models (Yuan and Bentler, 2007). Unfortunately, neither one of the two approaches can be used when parameters are constrained across levels—as is the case in the present study.

As an alternative, the Standardized Root Mean Square Residual (SRMR) can be calculated for both levels, which is equal to the square root of the squared standardized residual variances and covariances. The SRMR is computed as

$$\text{SRMR} = \sqrt{\frac{1}{t+p}\left(\sum_{k\leq l}(\hat{\varepsilon}^*_{kl})^2 + \sum_{k}(\hat{\varepsilon}^*_{l})^2\right)},$$

with

$$\hat{\varepsilon}^*_{kl} = \frac{s_{kl}}{\sqrt{s^2_k s^2_l}} - \frac{\hat{\sigma}_{kl}}{\sqrt{\hat{\sigma}^2_k \hat{\sigma}^2_l}} \text{ and } \hat{\varepsilon}^*_k = \frac{m_k}{\sqrt{s^2_k}} - \frac{\hat{\mu}_k}{\sqrt{\hat{\sigma}^2_k}},$$

where $t = \frac{p(p+1)}{2}$ is the number of (non-redundant) variances and covariances, s_{kl} denotes the sample covariance between variables k and l, s^2_k the sample variance of variable k, and s^2_l the sample variance of variable l. The model implied counterparts are $\hat{\sigma}_{kl}$, $\hat{\sigma}^2_k$, and $\hat{\sigma}^2_l$. Moreover, m_k and μ_k denote the sample and model implied mean of variable k. The SRMR is suitable for assessing how well the model in question reproduces the observed associations among the variables in an interpretable manner. With a grain of salt, it can be interpreted as the average of the absolute value of residual correlations. The SRMR can be calculated at both Level 1 and Level 2, thus offering a means to evaluate model fit within persons and between persons.[4] For the SRMR, a cut-off criterium of 0.08 has been recommended as based on simulation studies (Hu and Bentler, 1999).

All analyses reported below were conducted using Mplus, Version 7.11 Muthén and Muthén (2013). The absolute goodness-of-fit of models was evaluated using the Satorra-Bentler corrected χ^2-test and the Root Mean Square Error of Approximation (RMSEA). In addition, we report the Standardized Root Mean Square Residual (SRMR) for both the within- and the between-person covariance matrix. For both the RMSEA and the SRMR values <0.08 indicate acceptable model fit, whereas values <0.06 indicate good model fit (Hu and Bentler, 1999). In comparing the relative fit of nested models, we also detail the Satorra-Bentler corrected χ^2- difference test (Satorra and Bentler, 2010)—which, however, is expected to show excessive statistical power due to the large sample size. Thus, we based our decisions on which model to accept mainly on the SRMR within and between persons.

One additional remark appears in order here. While on Level 2 (between persons), the seven CES Items for the positive and negative events can covary, this is impossible at Level 1 (within persons), because an event is either positive or negative. As a consequence, while on Level 2 there are $\frac{14\times13}{2} = 91$ covariances, on Level 1 there are only $2 \times \frac{7\times6}{2} = 42$ covariances. To make such a model amenable for parameter estimation using MPLUS, the Level 1 covariance between the seven CES Items for the positive events and the seven CES Items for the negative events were constrained to be zero. At the same time, the total number of degrees of freedom was reduced by 49 in each model in order to achieve correct Satorra-Bentler corrected χ^2-tests and RMSEAs.

3 Results

Table 2 contains sample statistics for the seven CES items separately for positive and negative events. Shown are the sample means, within-person (Level 1) standard deviations, between-person (Level 2) standard deviations, and intraclass correlations. Two observations are key in Table 2 : (1) The intraclass correlations show that, in general, the amount of variance is smaller on Level 2 (between persons) than on Level 1 (within persons). In other words, participants differ more with respect to their CES ratings of the 10 positive and 10 negative events they evaluated than they differ from each other. (2) The intraclass correlations are, on average, lower for positive compared to negative events (0.266 vs. 0.358).[5]

3.1 Multilevel factor analyses

In a first model (Model 1 in Table 3), a one-factor model of centrality was estimated for both positive and negative events and at both levels of analysis (within and between persons) simultaneously.[6] As can be seen from the fit indexes listed in Table 3, Model 1 did not fit. An exploratory factor analysis indicated that a two-factor model (with Items 1, 2, 3, 4 loading on one factor and Items 5, 6, 7 loading on a second, correlated factor) described the data adequately.[7] Thus, in Model 1a, these two factors were specified within persons (Level 1), while between persons we continued with one factor. Although Model 1a represented a large improvement of fit compared to Model 1 (see Table 3), the RMSEA was not fully acceptable. Moreover, the $SRMR_b$ indicated that data were not described adequately on Level 2. For Model 1b, we "reversed" Model 1a by specifying one factor on Level 1 and two factors on Level 2. Although doing so also improved fit considerably compared to Model 1, the RMSEA was even less acceptable.

For Model 2, two factors were estimated at both levels of analysis and for both positive and negative events. This model (see Table 3) showed an acceptable fit. Moreover, it represented a huge improvement of fit compared to Model 1. Based on the factor loadings, we interpreted the first factor as capturing the impact of an event on a person's self-perception—in what follows, we abbreviate this Self-Perception factor as SP. More specifically, the factor captures the amount of which an event became integrated in one's life story and identity. The second factor, by contrast, can be interpreted as the impact of an event on one's (future) life-course—in what follows, we abbreviate this Life-Course factor as LC. Here, the consequences and implications of an event are in focus.[8] Along another dimension, one could also see the SP factor as capturing the inward-bound impact of an event on the self, requiring integration and reflection, whereas the LC gathers the outward-bound impact

4 Note that at Level 1, the part involving means is omitted.

5 The average within-person correlation among negative events was r = 0.549 (Median r = 0.535). The average within-person correlation among positive events was r = 0.572 (Median r = 0.554). Between persons, the average correlation was r = 0.615 (Median r = 0.606).

6 On both levels, factors were scaled by fixing the sum of their variances to 1. This scaling appears more adequate than the indicator-variable method, where the loading of one manifest indicator variable is fixed to 1, because in a measurement invariance analysis the latter would already implicitly assume equal factor loadings for this marker variable—something that goes untested until the weak measurement invariance model.

7 Table A1 in the Appendix contains some details on the tow-level exploratory factor models.

8 Note that on the SP factor items from all three theoretically postulated factors of centrality (Berntsen and Rubin, 2006) loaded: Item 1 (Identity), Item 2 (Identity), Item 3 (Turning Point), Item 4 (Reference Point). On the LC factor, items from two theoretically postulated factors loaded: Item 5 (Turning Point), Item 6 (Reference Point), Item 7 (Turning Point).

10.3389/fpsyg.2023.1268283

TABLE 2 Descriptive statistics of the CES items.

	Positive events (n = 2,712)				Negative events (n = 2,369)			
CES-Item	Mean	SD_w	SD_b	ICC	Mean	SD_w	SD_b	ICC
1. I feel that this event has become part of my identity.	3.637	1.142	0.682	0.263	3.195	1.118	0.862	0.373
2. This event has become a reference point for the way understand myself and the world.	3.161	1.136	0.759	0.309	3.057	1.093	0.857	0.382
3. I feel that this event has become a central part of my life story.	3.495	1.240	0.697	0.240	3.158	1.185	0.859	0.345
4. This event has colored the way I think and feel about other experiences.	2.964	1.133	0.746	0.303	3.199	1.077	0.834	0.375
5. This event permanently changed my life.	3.219	1.367	0.673	0.195	3.029	1.217	0.841	0.323
6. I often think about the effects this event will have on my future.	2.453	1.169	0.805	0.321	2.444	1.173	0.903	0.372
7. This event was a turning point in my life.	2.938	1.378	0.762	0.234	2.749	1.246	0.881	0.334

SD_w, within-person (level 1) standard deviation; SD_b, between-person (level 2) standard deviation; ICC, intraclass correlation.

TABLE 3 Model fit.

Model	χ^2_{SB}	*df*	SC	$\Delta\chi^2_{SB}$	Δdf	RMSEA	$SRMS_w$	$SRMR_b$
1: 1 factor at both levels	3,949*	104	0.793			0.085	0.044	0.101
1a: 2 factors within, 1 factor between	2,367*	102	0.785	1,060*[a]	2	0.066	0.028	0.082
1b: 1 factor within, 2 factors between	3,241*	99	0.818	1,612*[a]	5	0.079	0.043	0.062
2: 2 factors at both levels	1,962*	97	0.799	2,203*[a]	7	0.061	0.028	0.062
2a: 2 factors with residual covariances	454*	86	0.756	1,078*	11	0.029	0.014	0.039
3: 3 factors at both levels	3,208*	84	0.723	747*[a]	20	0.072	0.038	0.076
4: weak invariance I	447*	91	0.775	3[b]	5	0.028	0.014	0.039
5: weak invariance II	468*	96	0.778	21*	5	0.028	0.014	0.040
6: weak invariance III	623*	101	0.805	103*	5	0.031	0.013	0.057
7: strong invariance I	929*	106	0.811	270*	5	0.039	0.013	0.068
7a: strong invariance II	737*	107	0.809	263*	1	0.034	0.013	0.058
8: strict invariance I	735*	112	0.841	14	7	0.033	0.014	0.059
9: strict invariance II	739*	119	0.859	15	7	0.032	0.014	0.059
10: strict invariance III	1,958*	126	0.913	785*	7	0.054	0.015	0.260

$p < 0.01$, [a]compared to Model 1, [b]compared to Model 2a. χ^2_{SB}, Satorra-Bentler corrected chi-square; *df*, degrees of freedom; $\Delta\chi^2_{SB}$, difference in Satorra-Bentler corrected chi-square values; Δdf, difference in degrees of freedom; RMSEA, Root Mean Square Error of Approximation; $SRMS_w$, Standardized Root Mean Square Residual within persons (Level 1); $SRMS_b$, Standardized Root Mean Square Residual between persons (Level 2).

Model 4 = on the within-person level, factor loadings are constrained to be equal for positive and negative events, i.e., $\boldsymbol{\Lambda}_{wp} = \boldsymbol{\Lambda}_{wn} = \boldsymbol{\Lambda}_w$.

Model 5 = on the between-person level, factor loadings are constrained to be equal for positive and negative events, i.e., $\boldsymbol{\Lambda}_{bp} = \boldsymbol{\Lambda}_{bn} = \boldsymbol{\Lambda}_b$.

Model 6 = on both levels, factor loadings are constrained to be equal for positive and negative events, i.e., $\boldsymbol{\Lambda}_b = \boldsymbol{\Lambda}_w = \boldsymbol{\Lambda}$.

Model 7 = intercepts of CES items for positive and negative events are constrained to be equal, i.e., $\boldsymbol{\upsilon}_p = \boldsymbol{\upsilon}_n = \boldsymbol{\upsilon}$.

Model 7a = equality constraint of equal intercepts for positive and negative events relaxed for Item 4.

Model 8 = on the within-person level, residual variances are constrained to be equal for positive and negative events, i.e., $\boldsymbol{\Theta}_{ep} = \boldsymbol{\Theta}_{en} = \boldsymbol{\Theta}_e$.

Model 9 = Model 8 plus, on the between-person level, residual variances are constrained to be equal for positive and negative events, i.e., $\boldsymbol{\Theta}_{up} = \boldsymbol{\Theta}_{un} = \boldsymbol{\Theta}_u$.

Model 10 = Model 9 plus, on both levels, residual variances are constrained to be equal for positive and negative events, i.e., $\boldsymbol{\Theta}_u = \boldsymbol{\Theta}_e = \boldsymbol{\Theta}$.

of an event on a person's life, being aware of its implications and consequences. Factors were strongly correlated at both levels of analysis and for both positive and negative events (*r*s ranging from 0.52 to 0.88). The model is depicted in Figure 1.

Because according to the RMSEA, fit was at the boundary of the typical cut-off (0.06) of good model fit, in the next model (Model 2a in Table 3), we introduced covariances between the same respective items for positive and negative items on Level 2, the between-person level (i.e., between Item 1 for positive events and Item 1 for negative events, etc.).[9] Moreover, on Level 1 (within persons), we introduced residual covariances between Items 1 and 2 for positive and negative events and for Items 2 and 4 for positive and negative events.[10] This model (Model 2a) showed an improved fit, which, in addition, represented an improvement compared to the previous model.

For reasons of completeness, we also estimated a three-factor model with the seven items designated to load on their respective theoretically proposed factors. As can be seen from Table 3, this model (Model 3) did not describe the data well. Furthermore, factors virtually collapsed, that is, their correlations approached unity. Therefore, we decided to continue with Model 2a, which served as the configural invariance model for the measurement invariance analyses.

3.2 Measurement invariance analyses

In examining measurement invariance, in a first model (Model 4 in Table 3), we imposed weak invariance with respect to positive and negative events at the within-person level (i.e., $\boldsymbol{\Lambda}_{wp} = \boldsymbol{\Lambda}_{wn} = \boldsymbol{\Lambda}_{w}$). This model showed an acceptable fit, which, moreover, was indistinguishable from that of Model 2a. Based on this result, we concluded that weak MI holds for measuring event centrality for different events (positive vs. negative) within persons.

In the next model (Model 5), the constraint of equal factor loadings for positive and negative events between persons was added (i.e., $\boldsymbol{\Lambda}_{bp} = \boldsymbol{\Lambda}_{bn} = \boldsymbol{\Lambda}_{b}$). Although the Satorra-Bentler corrected χ^2-difference indicated a significant loss of fit, the RMSEA and both SRMRs remained virtually unchanged, from which we inferred that weak MI holds for measuring event centrality across different events (positive vs. negative) between persons.

Model 6 imposed equal factor loadings across event type and across levels, thus implying "complete" weak measurement invariance (i.e., $\boldsymbol{\Lambda}_{w} = \boldsymbol{\Lambda}_{p} = \boldsymbol{\Lambda}$). As Table 3 shows, doing so led to a relatively large decrement in model fit. At the same time, the RMSEA and both SRMRs were still well below their critical cut-off criterium. For this reason, we regarded Model 6 as adequately describing the data.

In Model 7, latent intercepts of the CES items were constrained to be equal across positive and negative events (i.e., $\boldsymbol{\upsilon}_{p} = \boldsymbol{\upsilon}_{n} = \boldsymbol{\upsilon}$). From the fit indexes in Table 3, it becomes apparent that this led to a relatively large decrease in fit on Level 2 (in line with the fact that latent intercepts constraints only affect the between-person data level). Upon inspection, Item 4 ("This event has colored the way I think and feel about other experiences.") showed a large discrepancy for positive and negative events. Therefore, in Model 7a, the constraint of equal intercepts for positive and negative events was relaxed for Item 4. This model showed an almost unchanged fit compared to Model 6. Results showed that Item 4 was endorsed more strongly for negative events than what would have been expected based on the Self-Perception factor differences, while it was endorsed less strongly for positive events. Taken together, only partial strong measurement invariance held across positive vs. negative events.

In Model 8, residual variances were constrained to be equal for corresponding CES items for positive and negative events at the within-person level ($\boldsymbol{\Theta}_{ep} = \boldsymbol{\Theta}_{en} = \boldsymbol{\Theta}_{e}$). As can be seen from Table 3, doing so left model fit almost unchanged. Next, for Model 9, the constraint of equal residual variances for corresponding CES items for positive and negative events at the between-person level was added to Model 8 ($\boldsymbol{\Theta}_{up} = \boldsymbol{\Theta}_{un} = \boldsymbol{\Theta}_{u}$). Again, model fit remained virtually the same. Finally, in Model 10, residual variances were, in addition to Model 9, required to be equal within and between persons ($\boldsymbol{\Theta}_{u} = \boldsymbol{\Theta}_{e} = \boldsymbol{\Theta}$). As expected, this model did not achieve an adequate fit.

Summarizing these analyses, we accepted Model 9 as reflecting the associations among CES items for positive and negative events *and* within and between persons adequately. Model 9 entails the following elements of measurement invariance: (1) Factor loadings are completely equal, that is, $\boldsymbol{\Lambda}_{wp} = \boldsymbol{\Lambda}_{wn} = \boldsymbol{\Lambda}_{bp} = \boldsymbol{\Lambda}_{bn} = \boldsymbol{\Lambda}$. This implies that factor variances and covariances can be compared across event types and across data levels. Figure 2 depicts the factor variance estimates based on Model 9. If the 84% inferential confidence intervals (see Tryon, 2001) of (any) two factor variances do not overlap, the variances are significantly different from each other ($p < 0.05$). In line with the descriptive statistics (see Table 2), factor variances were larger on Level 1 (within persons) than on Level 2 (between persons). Moreover, on both levels, the factor variances of Self-Perception were larger than for Life-Course, implying that both event differences and individual differences were more pronounced for Self-Perception than for Life-Course. In addition, Figure 3 shows the factor covariance between Self-Perception and Life-Course for positive and negative AMs and on both data levels. As for the factor variances, factor covariances are much larger on Level 1, implying that the centrality assessments – Self-Perception and Life-Course – are more similar within persons than between persons.[11] (2) Item intercepts are equal for positive and negative events (except Item 4), implying

9 These residual covariances appear justified based on the assumption that there is an individual, idiosyncratic tendency to rate the respective positive and negative CES items similarly, e.g., generally endorsing Item 1 strongly for both positive and negative events.

10 The residual covariance between Items 1 and 2 is most likely due to both items belonging to the "identity" factor of the full CES. For Items 2 and 4, there is no obvious reason for a residual covariance.

11 Note that factor correlations (or standardized covariances) are much more similar across levels (see Figure 1), which results from the fact that factor variances were also much larger on Level 1. For a comparison of the strength of relationships among factors across levels, however, covariances represent the more adequate metric because correlations are based on the assumption of equal variances—which obviously does not hold (see Figure 2).

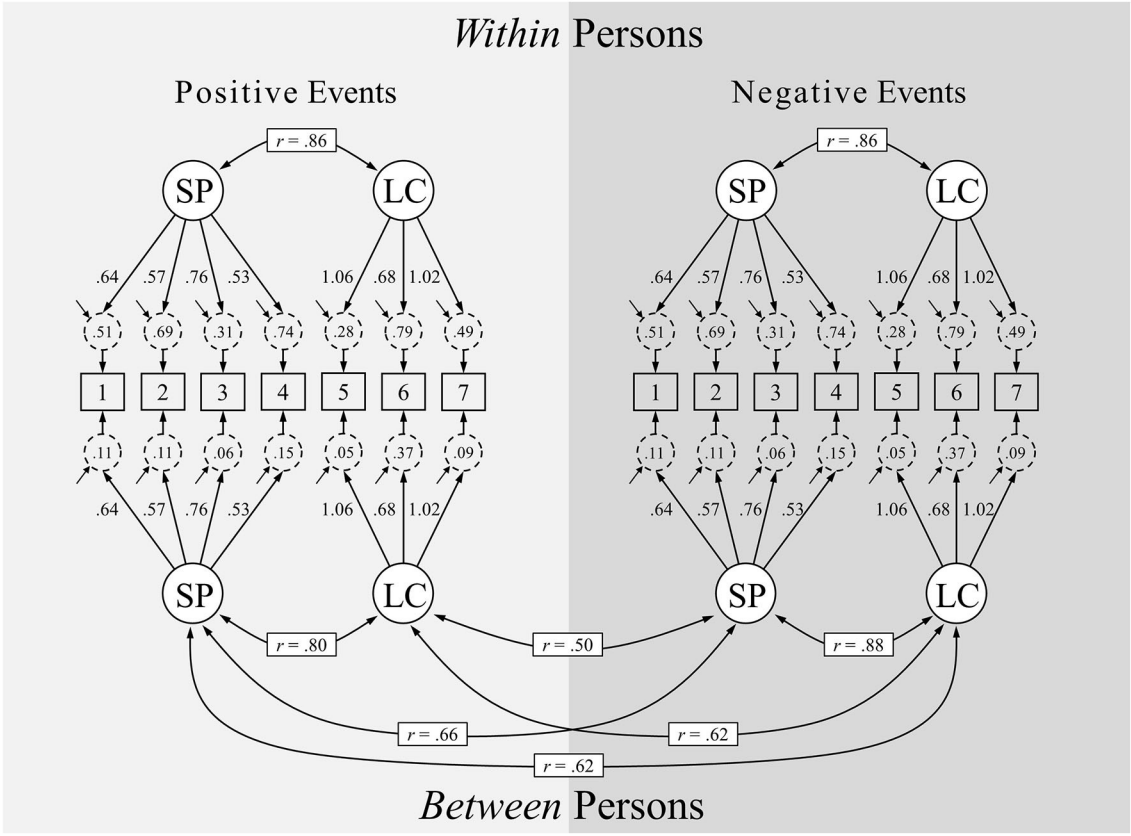


FIGURE 1
Multilevel factor analysis model of the short version of the Centrality of Event Scale (based on Model 9). SP, impact of an event on a person's Self-Perception; LC, impact of an event on a persons's (future) Life-Course. Apart from inter-factor correlations, parameters are unstandardized.

partial strong invariance. Based on Model 9, factor means can be compared on Level 2 (keeping in mind that factor means were modeled without Item 4).[12] Figure 4 shows the according factor means scaled in the effect size metric of Cohen's *d*—note that the factor means of the negative events were constrained to be zero for identification purposes, such that the factor means of the positive events represent the difference. The (factor) mean difference of Self-Perception between negative and positive events amounted to an effect size of 0.58, which conventionally would be regarded a medium effect. Thus, the impact of positive events on one's Self-Perception was judged as larger than that of negative events. By contrast, for the difference in Life-Course between negative and positive events, the effect size was 0.32, a small effect. Hence, the impact of positive events on one's Life-Course was larger than that of negative events—although the effect was only about half of the size of the Self-Perception effect. (3) Residual variances of the seven CES items were equal for positive and negative at the within-person and the between-person level, but not across levels. This implies that conditional variance of item responses (given the SP and LC factors), is invariant for intraindividual differences between positive and negative events and interindividual differences.

4 Discussion

A person's life story is composed of personally experienced events that are considered highly self-relevant at the time when they took place or which maintain self-importance over time (Bluck and Habermas, 2000; Conway and Holmes, 2004). The life story provides a person with an overall sense of meaningfulness, purpose, and coherence (McAdams, 2001), and thus fosters a sense of self-identity (Conway and Tagini, 2004). However, not all personally experienced events become part of a person's life story; and even those that do, may vary in terms of their self-relevance. For instance, people typically consider their most positive autobiographical memory as more central to their identity than their most negative one (e.g., Zaragoza Scherman et al., 2020; Pociūnaitė and Zimprich, 2023). The centrality of an event may not only vary as a function of valence (i.e., positive vs. negative memories), but also within valence categories in the sense that some positive (or negative) memories contribute strongly to a person's identity and life story, whereas other positive (or negative) events are perceived as less self-relevant. Against this background, it is important to ensure that self-report questionnaires tapping

12 Only for the Self-Perception factor strong invariance was partial (because Item 4 is an indicator of it), while for Life-Course full strong invariance held across event type.

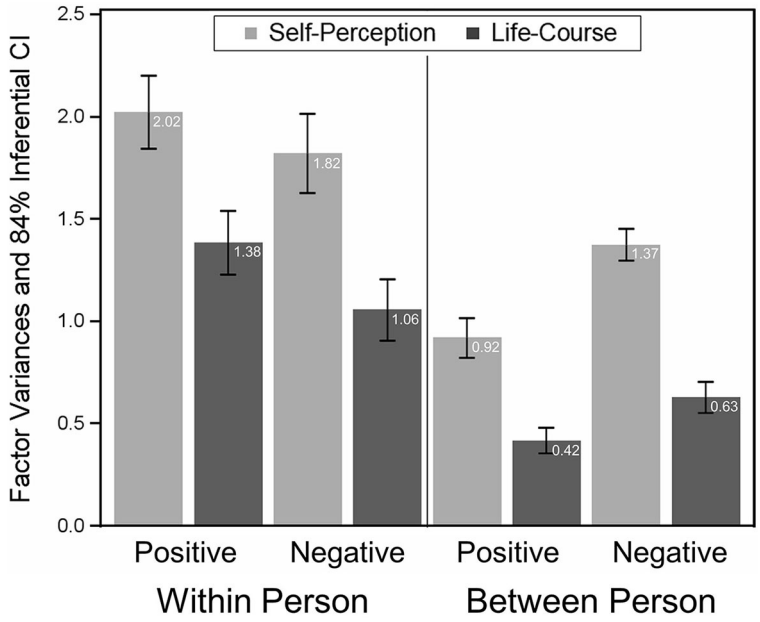


FIGURE 2
Factor variances of self-perception and life-course for positive and negative events *and* within and between persons. Also shown are the 84% inferential confidence intervals (see Tryon, 2001).

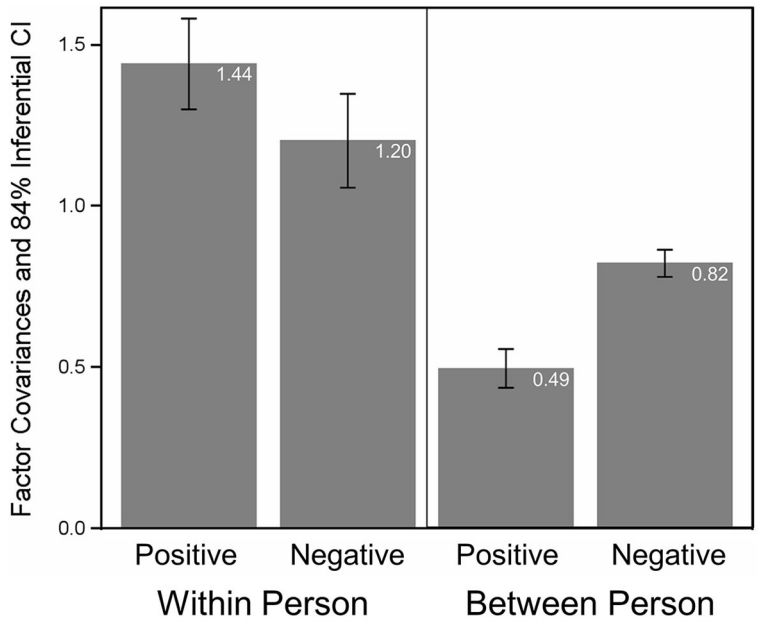


FIGURE 3
Factor covariances of self-perception and life-course for positive and negative events *and* within and between persons. Also shown are the 84% inferential confidence intervals (see Tryon, 2001).

the degree to which autobiographical memories are embedded in a person's life story are reliable measures on both the between-person level as well as the within-person level. The present study provides a first examination of the seven items included in the Centrality of Event Scale (CES) short form. Based on exploratory and confirmatory factor analyses, we found a two-factor structure (Self-Perception and Life Course) at both levels of analyses *and* for positive and negative events.

4.1 One or two factors of event centrality?

A few studies have tested the factorial structure of the CES short form and they univocally advocate for a one-factor solution. Given that the short form consists of seven items only, a one-factor solution seems both plausible and practical. Depending on the research question, however, a more fine-graded measure seems warranted; for instance, to understand why some events

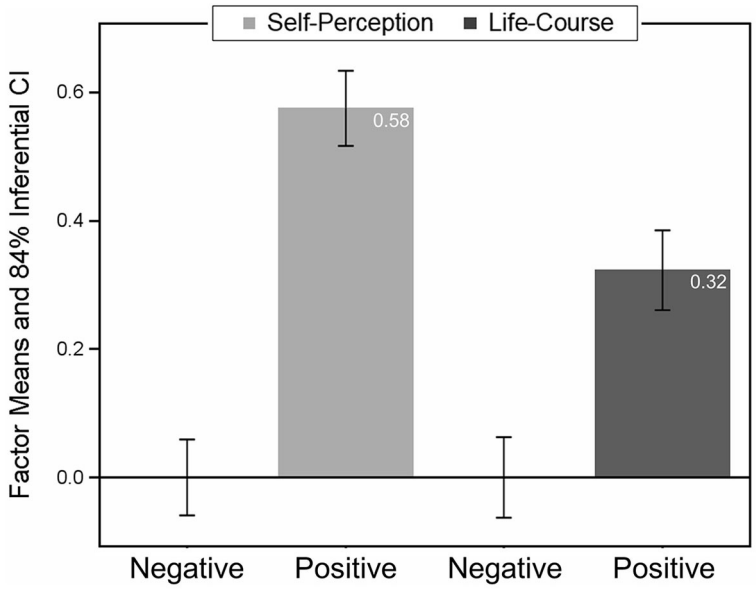


FIGURE 4
Between-person factor means of self-perception and life-course for positive and negative events (level 2). Also shown are the 84% inferential confidence intervals (see Tryon, 2001).

are perceived as more central than other events. Much like the full version of the CES (e.g., Ionio et al., 2018), the short from is comprised of three, theoretically distinct components: Events can form a central component of personal identity (two items), a turning point in the life story (three items), and a reference point for everyday inferences (two items). These three components are not that clearly mirrored on our two-factor solution. In fact, the first factor included items from all three theoretically postulated factors. What these items share is a focus on the impact of an event on a person's self-perception in the sense that the event shapes how the person thinks and feels about themselves. The first factor, thus, captures the inward-bound impact of an event on the cognitive and emotional level, which requires reflecting on the event and integrating it into one's identity and life story. The second factor consists of items capturing the degree to which an event represents a turning, or reference point; thus only capturing two of the theoretically proposed factors. What these three items share is a focus on the event's impact on a person's life-course—be it in the past or anticipated in the future. Put differently, the second factor describes the outward-bound impact of an event, respectively, the implications and consequences of an event for a person's life.

Notably, these two factors show different patterns in terms of factor variances and covariances on the two data levels: Variances and covariances were much more pronounced within persons than between persons. This implies that individuals differ in assessing the amount of Self-Perception and Life-Course of their individual autobiographical memories (Level 1 variances) but are much more similar when all autobiographical memories are considered together (Level 2 variances). This also shows in the factor covariances, where Self-Perception and Life-Course centrality assessments are more strongly related on Level 1 than on Level 2. In sum, this indicates that there are individuals who tend to go to more extremes in assessing individual positive or negative autobiographical memories, but across all reported autobiographical memories these extremes become more equalized such that individuals are more similar. With respect to factor means (between-person level only), participants generally perceived their positive memories as more central than their negative memories—as indicated by higher factor means for positive compared to negative memories—but this difference was more pronounced for the factor capturing the impact of an event on a person's self-perception (medium effect) compared to the factor describing an event's impact on the life-course (small effect). This implies that both positive and negative events have the potential to change a person's life, be it in a positive or negative way. However, they show distinct contributions to a person's self-perception in the sense that positive events, in particular, shape how a person is thinking and feeling about themselves, their identity, and the world. This aligns with research showing that positive and negative memories serve different functions in daily life (Rasmussen and Berntsen, 2009). For instance, positive memories are more often used to feel better (about oneself), whereas negative memories serve to direct behavior with the goal of avoiding similar experiences, and their negative impact upon one's life in the future (Wolf and Demiray, 2019; Wolf et al., 2021).

Our factorial structure of the short, seven-item version of the CES differs from that found in previous research. Whereas, in previous studies, the short version typically evinced one underlying factor, we found two, albeit substantially correlated factors on the between-person level both for positive and negative events. There are several possible reasons for this discrepancy. First, in our study not traumatic, but simply positive and negative events from their past were assessed using the brief CES. Therefore, one would expect that (a) centrality in our study is, in general, lower

than for high-impact, traumatic events and (b) that centrality is more variable across events. This may have led to lower inter-item correlations on both levels of analysis compared to previous studies. Second, between-person differences in our study were not based on having every participant evaluate one event, but result as the individual-specific means across CES items across up to 10 positive and 10 negative events. Associations on Level 2 are expected to be different among items (as analyzed in previous studies) vs. among (latent) person-specific means of items (as analyzed in our study). Importantly, reliability of individual differences can be assumed to be higher in the approach we used (e.g., Muthén, 1991).

To understand the different roles of positive and negative events for a person's identity and life story, a more nuanced centrality measure seems to offer a more fine-graded picture. This does not necessarily imply that a two-factor solution needs to always be applied when using the CES short form. A unidimensional scale might be sufficient when focusing on a person's most stressful or traumatic life event (e.g., Galán et al., 2017; Vagos et al., 2018; Vermeulen et al., 2020; Azadfar et al., 2022), because for highly stressful or traumatic events, one would expect a relatively similar, strong endorsement of all CES items.

4.2 Measurement invariance of the brief CES

In the present study—to the best of our knowledge for the first time—the measurement properties of the brief CES were examined both between and within persons and, simultaneously, for positive and negative events (see Figure 1). The model we accepted (Model 9) shows that factor loadings were completely invariant across the quadrants of the scheme in Table 1. That is, weak measurement invariance was established, which allows for a direct comparison of factor variances across event types (positive vs. negative) and across levels of analysis. From Figure 2, it becomes evident that factor variances of Self-Perception and Life-Course were, in general, larger within persons than between persons, implying larger differences among events than among individuals. Moreover, variances of Self-Perception were larger than variances of Life-Course, indicating that the amount of which events become integrated in one's life story and identity (the inward-bound effect of events) was more variable than the amount of which an event has implications and consequences for one's life (the outward-bound effect of events). This appears to suggest that a person can have varying internal interpretations of an event, whereas the external implications is more objective or more universal.

Intercepts of the CES items were not fully invariant across event types because Item 4 showed a pattern different from the remaining items. Whereas, for the other items, both Self-Perception and Life-Course were more pronounced for positive events, amounting to a medium and a small effect (see Figure 4), for Item 4 this pattern was reversed. Thus, negative events appear to color the way individuals think (and feel) about other experiences more than positive events do. This finding has a potentially important consequence: A comparison of the centrality of positive and negative events might better exclude Item 4, because it (with its reverse effect) leads to a downward bias of the event centrality difference. As such, one might suspect that the centrality differences between positive and negative events reported in the literature (e.g., Zaragoza Scherman et al., 2015) may underestimate the true difference.

One implication of weak invariance holding across levels concerns the definition of the intraclass correlation coefficient. Muthén (1991) proposed a "true" intraclass correlation coefficient (ρ_{icc}), which makes use of the factor-analytic decomposition of the observed variance into a systematic and a residual part and gives the error-free proportion of between-person variance (see Equation 1). For variables that load on one factor only (congeneric model)—as in the present analysis of the brief CES—we have

$$\rho_{icc} = \frac{\lambda_b^2 \phi}{\lambda_b^2 \phi + \lambda_w^2 \psi}, \tag{2}$$

where λ_b is the factor loading of the item in question on Level 2, ϕ is the variance of the factor on Level 2, λ_w is the factor loading on Level 1, and ψ is the variance of the factor on Level 1. By contrast to the ordinary intraclass correlation coefficient, this "true" intraclass correlation coefficient is not contaminated by measurement error. At the same time, however, it is a model-based quantity based on factor variances, which may take on different values depending on the model used to estimate it. Based on our Model 9 with equal factor loadings on the within-person and the between-person level, Equation (2) can be further simplified, such that (cf. Zimprich and Martin, 2009)

$$\rho_{icc} = \frac{\lambda_b^2 \phi}{\lambda_b^2 \phi + \lambda_w^2 \psi} = \frac{\lambda^2 \phi}{\lambda^2 (\phi + \psi)} = \frac{\phi}{\phi + \psi},$$

implying equal "true" intraclass correlation coefficients for those items loading on the same factor (SP vs. LC). From a substantive perspective equal "true" intraclass correlations appear reasonable: Those variables measuring the same underlying factor have the same ratio of "true" between-person variance in comparison to the total "true" variance—with this ratio being independent of the actual scaling of variables.

5 Conclusion

The Centrality of Event Scale (CES) was originally developed to measure the extent of which a traumatic or stressful event becomes integrated into a person's identity and life story. Our findings demonstrate that the CES constitutes a reliable measure to compare the centrality of emotionally positive and emotionally negative memories within and between persons. How the CES is analyzed, however, may depend on the type of events researchers are focusing on. When focusing on traumatic or highly stressful life events, the items of the CES short form may form a single factor. In aiming to understand the different roles of positive and negative events for a person's identity and life story, however, it seems warranted to distinguish between an event's impact on a person's Self-Perception and its consequences for a person's Life-Course. Moreover, the seven items of the CES short form may not be equally suited to meaningfully compare the centrality of positive and negative events (i.e., Item 4). Finally, based on the two-level interpretation, an event can have stronger influences on individual differences in self-perception, whereas the life-course-changing properties of events appear to be less variable across persons.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical approval was not required for the studies involving humans because at the time the study was conducted, ethical approval was not required at Ulm University. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

DZ: Formal analysis, Investigation, Methodology, Visualization, Writing–original draft, Writing–review & editing. JP: Formal analysis, Methodology, Visualization, Writing–review & editing. TW: Conceptualization, Data curation, Investigation, Project administration, Writing–original draft, Writing–review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Azadfar, Z., Khosravi, Z., Farah Bijari, A., and Abdollahi, A. (2022). The persian version of the centrality of event scale (ces): assessment of validity and reliability among iranian university students. *Brain Behav.* 12, e2448. doi: 10.1002/brb3.2448

Berntsen, D., and Rubin, D. C. (2006). The centrality of event scale: a measure of integrating a trauma into one's identity and its relation to post-traumatic stress disorder symptoms. *Behav. Res. Ther.* 44, 219–231. doi: 10.1016/j.brat.2005.01.009

Berntsen, D., Rubin, D. C., and Siegler, I. C. (2011). Two versions of life: emotionally negative and positive life events have different roles in the organization of life story and identity. *Emotion* 11, 1190–1201. doi: 10.1037/a0024940

Bluck, S., and Habermas, T. (2000). The life story schema. *Motiv. Emot.* 24, 121–147. doi: 10.1023/A:1005615331901

Boals, A. (2010). Events that have become central to identity: gender differences in the centrality of events scale for positive and negative events. *Appl. Cogn. Psychol.* 24, 107–121. doi: 10.1002/acp.1548

Boelen, P. A. (2012). A prospective examination of the association between the centrality of a loss and post-loss psychopathology. *J. Affect. Disord.* 137, 117–124. doi: 10.1016/j.jad.2011.12.004

Brown, A. D., Antonius, D., Kramer, M., Root, J. C., and Hirst, W. (2010). Trauma centrality and ptsd in veterans returning from iraq and afghanistan. *J. Trauma. Stress* 23, 496–499. doi: 10.1002/jts.20547

Bruce, M. J., and Handal, P. (2023). Revisiting the factor structure of the centrality of event scale. *OMEGA* 7, 00302228231162211. doi: 10.1177/00302228231162211

Conen, L., Johanßen, H., Ülsmann, D., Ertle, A., Schulte, S., Fydrich, T., et al. (2022). Validierung der deutschen Übersetzung der centrality of event scale (ces-g). *Zeitschrift Klin. Psychol. Psychother.* 51, 47–55. doi: 10.1026/1616-3443/a000651

Conway, M. A., and Holmes, A. (2004). Psychosocial stages and the accessibility of autobiographical memories across the life cycle. *J. Person.* 72, 461–480. doi: 10.1111/j.0022-3506.2004.00269.x

Conway, M. A., Singer, J. A., and Tagini, A. (2004). The self and autobiographical memory: correspondence and coherence. *Soc. Cogn.* 22, 491–529. doi: 10.1521/soco.22.5.491.50768

Galán, S., Castarlenas, E., Racine, M., Sánchez-Rodríguez, E., Tomé-Pires, C., Jensen, M. P., et al. (2017). Factor structure, internal consistency and criterion calidity of the full-form and short-form versions of the centrality of events scale in young people. *Appl. Cogn. Psychol.* 31, 662–667. doi: 10.1002/acp.3369

Gauer, G., Souza, J. A., Silveira, A. M., and Sediyama, C. Y. N. (2013). Stressful events in autobiographical memory processing: Brazilian version of the centrality of event scale. *Psicologia* 26, 98–105. doi: 10.1590/S0102-79722013000100011

Groleau, J. M., Calhoun, L. G., Cann, A., and Tedeschi, R. G. (2013). The role of centrality of events in posttraumatic distress and posttraumatic growth. *Psychol. Trauma* 5, 477–483. doi: 10.1037/a0028809

Habermas, T. (2019). *Emotion and Narrative: Perspectives in Autobiographical Storytelling*. Cambridge: Cambridge University Press.

Hoehne, S. (2023). Can perceived changes in autobiographical memories' emotionality be explained by memory characteristics and individual differences? *Memory* 31, 850–863. doi: 10.1080/09658211.2023.2207803

Hox, J. (1995). *Applied Multilevel Analysis*. Amsterdam: TT Publikaties.

Hu, L.-T., 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

Ionio, C., Mascheronia, E., and Di Blasio, P. (2018). The centrality of events scale for italian adolescents: Integrating traumatic experience into one's identity and its relation to posttraumatic stress disorder symptomatology. *Europes J. Psychol.* 14, 359–372. doi: 10.5964/ejop.v14i2.1465

Leiner, D. J. (2019). *SoSci Survey (Version 3.1.06)* [*Computer software*]. Available online at: https://www.soscisurvey.de

Longford, N. T., and Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika* 57, 581–597. doi: 10.1007/BF02294421

McAdams, D. P. (2001). The psychology of life stories. *Rev. Gen. Psychol.* 5, 100–122. doi: 10.1037/1089-2680.5.2.100

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika* 58, 575–585. doi: 10.1007/BF02294828

Mehta, P. D., and Neale, M. C. (2005). People are variables too: multilevel structural equation modeling. *Psychol. Methods* 58, 259–284. doi: 10.1037/1082-989X.10.3.259

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Meredith, W., and Horn, J. (2001). "The role of factorial invariance in modeling growth and change," in *New Methods for the Analysis of Change*, eds L. M. Collins, and A. G. Sayer (Washington, DC: American Psychological Association).

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *J. Educ. Meas.* 28, 338–354. doi: 10.1111/j.1745-3984.1991.tb00363.x

Muthén, L. K., and Muthén, B. O. (2013). *Mplus User's Guide*. 7th *Edn*. Los Angeles, CA: Muthn & Muthn.

Newby, J. M., and Moulds, M. L. (2011). Intrusive memories of negative events in depression: Is the centrality of the event important? *J. Behav. Ther. Exp. Psychiatry* 42, 277–283. doi: 10.1016/j.jbtep.2010.12.011

Pillemer, D. B. (2009). Twenty years after baddeley, Is the study of autobiographical memory fully functional? *Appl. Cogn. Psychol*. 23, 1193–1208. doi: 10.1002/acp.1619

Pinto-Gouveia, J., and Matos, M. (2011). Can shame memories become a key to identity? The centrality of shame memories predicts psychopathology. *Appl. Cogn. Psychol*. 25, 281–290. doi: 10.1002/acp.1689

Pociūnaitė, J., and Zimprich, D. (2023). Characteristics of positive and negative autobiographical memories central to identity: emotionality, vividness, rehearsal, rumination, and reflection. *Front. Psychol*. 14, 1225068. doi: 10.3389/fpsyg.2023.1225068

Pociūnaitė, J., Zimprich, D., and Wolf, T. (2022). Centrality of positive and negative autobiographical memories across the adult life span. *Appl. Cogn. Psychol*. 36, 623–635. doi: 10.1002/acp.3949

Rasmussen, A. S., and Berntsen, D. (2009). Emotional valence and the functions of autobiographical memories: positive and negative memories serve different functions. *Mem. Cogn*. 37, 477–492. doi: 10.3758/MC.37.4.477

Rasmussen, A. S., and Berntsen, D. (2013). The reality of the past versus the ideality of the future: emotional valence and functional differences between past and future mental time travel. *Mem. Cogn*. 41, 187–200. doi: 10.3758/s13421-012-0260-y

Rubin, D. C., Boals, A., and Hoyle, R. H. (2014). Narrative centrality and negative affectivity: independent and interactive contributors to stress reactions. *J. Exp. Psychol*. 143, 1159–1170. doi: 10.1037/a0035140

Ryu, E., and West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Struct. Equ. Model.* 16, 583–601. doi: 10.1080/10705510903203466

Satorra, A., and Bentler, P. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika* 75, 243–248. doi: 10.1007/s11336-009-9135-y

Schuettler, D., and Boals, A. (2011). The path to posttraumatic growth versus ptsd: contributions of event centrality and coping. *J. Loss Trauma* 16, 180–194. doi: 10.1080/15325024.2010.519273

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychol. Methods* 6, 371–386. doi: 10.1037/1082-989X.6.4.371

Vagos, P., Da Silva, D. R., Brazão, N., and Rijo, D. (2018). The centrality of event scale in portuguese adolescents: Validity evidence based on internal structure and on relations to other variables. *Assessment* 25, 527–538. doi: 10.1177/1073191116651137

Vermeulen, M., Smits, D., Boelen, P. A., Claes, L., Raes, F., and Krans, J. (2020). The dutch version of the centrality of event scale (ces). *Eur. J. Psychol. Assess*. 36, 361–371. doi: 10.1027/1015-5759/a000517

Walker, W. R., Skowronski, J. J., and Thompson, C. P. (2003). Life is pleasant–and memory helps to keep it that way! *Rev. Gen. Psychol*. 7, 203–210. doi: 10.1037/1089-2680.7.2.203

Walker, W. R., Vogl, R. J., and Thompson, C. P. (1997). Autobiographical memory: unpleasantness fades faster than pleasantness over time. *Appl. Cogn. Psychol*. 11, 399–413. doi: 10.1002/(SICI)1099-0720(199710)11:5<399::AID-ACP462>3.0.CO;2-E

Wamser-Nanney, R. (2019). Event centrality: factor structure and links to posttraumatic stress disorder symptom clusters. *J. Trauma. Stress* 32, 516–525. doi: 10.1002/jts.22413

Wolf, T., and Demiray, B. (2019). The mood-enhancement function of autobiographical memories: comparisons with other functions in terms of emotional valence. *Conscious. Cogn*. 70, 88–100. doi: 10.1016/j.concog.2019.03.002

Wolf, T., Pociūnaitė, J., Hoehne, S., and Zimprich, D. (2021). The valence and the functions of autobiographical memories: does intensity matter? *Conscious. Cogn*. 91, 103119. doi: 10.1016/j.concog.2021.103119

Wolf, T., and Zimprich, D. (2015). Differences in the use of autobiographical memory across the adult lifespan. *Memory* 23, 1238–1254. doi: 10.1080/09658211.2014.971815

Yuan, K.-H., and Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models. *Sociol. Methodol*. 37, 53–82. doi: 10.1111/j.1467-9531.2007.00182.x

Zaragoza Scherman, A., Salgado, S., Shao, Z., and Berntsen, D. (2015). Event centrality of positive and negative autobiographical memories to identity and life story across cultures. *Memory* 23, 1152–1171. doi: 10.1080/09658211.2014.962997

Zaragoza Scherman, A., Salgado, S., Shao, Z., and Berntsen, D. (2020). Younger adults report more distress and less well-being: a cross-cultural study of event centrality, depression, post-traumatic stress disorder and life satisfaction. *Appl. Cogn. Psychol*. 34, 1180–1196. doi: 10.1002/acp.3707

Zimprich, D., Allemand, M., and Hornung, R. (2006). Measurement invariance of the abridged sense of coherence scale in adolescents. *Eur. J. Psychol. Assess*. 22, 280–287. doi: 10.1027/1015-5759.22.4.280

Zimprich, D., Allemand, M., and Lachman, M. E. (2012). Factorial structure and age-related psychometrics of the midus personality adjective items across the life span. *Psychol. Assess*. 24, 173–186. doi: 10.1037/a0025265

Zimprich, D., and Martin, M. (2009). "A multilevel factor analysis perspective on intellectual development in old age," in *Aging and Cognition: Research Methodologies and Empirical Advances*, eds H. B, Bosworth, and C. Hertzog (Washington, DC: American Psychological Association).

Zimprich, D., Perren, S., and Hornung, R. (2005). A two-level confirmatory factor analysis of a modified rosenberg self-esteem scale. *Educ. Psychol. Meas*. 65, 465–481. doi: 10.1177/0013164404272487

10.3389/fpsyg.2023.1268283

Appendix

TABLE A1 Factor loadings in multilevel exploratory factor analyses of the brief CES.

	One-factor model		Two-factor model			
	Within	Between	Within		Between	
Item	Factor 1	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2
	Positive events					
1	0.76	0.88	0.52	0.35	0.95	0.00
2	0.69	0.92	0.89	0.00	0.76	0.26
3	0.86	0.85	0.55	0.38	0.68	0.32
4	0.66	0.93	0.64	0.14	0.54	0.26
5	0.89	0.76	0.00	0.93	0.08	0.89
6	0.69	0.40	0.09	0.62	−0.08	0.77
7	0.84	0.67	0.00	0.87	0.00	0.93
	Negative events					
1	0.75	0.89	0.46	0.33	0.89	0.00
2	0.70	0.93	0.88	0.00	0.94	−0.02
3	0.84	0.95	0.53	0.34	0.57	0.32
4	0.67	0.88	0.60	0.18	0.91	0.00
5	0.84	0.91	0.00	0.91	0.00	0.91
6	0.60	0.69	0.12	0.52	0.15	0.62
7	0.79	0.86	0.00	0.83	0.00	0.94

Geomin rotation was used in the two-factor model. Factor loadings were estimated using Maximum Likelihood.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 11 January 2024
DOI 10.3389/fpsyg.2023.1346232

OPEN ACCESS

EDITED BY
Georgios Sideridis,
Harvard Medical School, United States

REVIEWED BY
Dimitrios Stamovlasis,
Aristotle University of Thessaloniki, Greece
Angeliki Mouzaki,
University of Crete, Greece
Celestino Rodríguez,
Universidad de Oviedo Mieres, Spain

*CORRESPONDENCE
Ghadah Alkhadim
ghadah.s@tu.edu.sa

RECEIVED 29 November 2023
ACCEPTED 14 December 2023
PUBLISHED 11 January 2024

CITATION
Alkhadim G (2024) The detrimental effects of student-disordered behavior at school: evidence from using the cusp catastrophe. *Front. Psychol.* 14:1346232. doi: 10.3389/fpsyg.2023.1346232

The detrimental effects of student-disordered behavior at school: evidence from using the cusp catastrophe

Ghadah Alkhadim*

Faculty of Educational Psychology, Taif University, Taif, Saudi Arabia

Introduction: The purpose of the present study is to examine the potentially complex relationship between disordered behavior at school and students' engagement with reading activities given that they enjoy reading. Of particular interest is the role of disordered behavior which we believe moderated the relationship between liking reading and reading engagement.

Methods: Participants were 2,420 fourth graders who participated in the 2021 PIRLS study from Saudi Arabia and were selected using stratified random sampling from 117 schools in the Kingdom. Data were analyzed using linear and nonlinear means such as the linear model, the logistic model, and the cusp catastrophe.

Results: Results pointed to the superiority of the cusp catastrophe towards predicting student engagement in reading by highlighting the splitting role of students' disruptive classroom behavior.

Discussion: It was evident that exceeding a critical upward level in disruptive classroom behavior was associated with unpredictable and sudden changes in reading engagement. It is concluded that the application of non-linear means may be conducive to understanding complex educational phenomena.

KEYWORDS

reading engagement, liking of reading, disordered behavior, cusp catastrophe, nonlinear modeling, nonlinear dynamics systems theory

1 Introduction

Students' attitudes about reading are significantly impacted by the concept of engagement in reading, which includes behavioral, emotional, and cognitive components (Fredricks et al., 2004). According to Guthrie and Klauda (2008), this concept is a crucial factor in shaping students' attitudes toward reading. When it comes to reading, behavioral engagement refers to the active participation in reading activities, whereas emotional engagement comprises the subjective emotional reactions that are experienced when reading such as joy and intrinsic interest (Baker et al., 2000; Unrau and Schlackman, 2006). On the other side, cognitive engagement refers to the mental effort and investment that is made to master the content that is being read (Fredricks et al., 2004) and involves the use of deep processing and strategic reading behaviors such as previewing, visualizing, monitoring, making connections, synthesizing, and summarizing (Pressley and Afflerbach, 1995; Wigfield and Guthrie, 1997; Duke and Pearson, 2002; Afflerbach et al., 2008).

The scientific literature emphasizes the importance of reading enjoyment in influencing students' engagement with reading (Martínez et al., 2008; Smith et al., 2012; Afflerbach et al., 2013; Mol and Jolles, 2014; Lim et al., 2015; Ho and Lau, 2018; Merga and Roni, 2018; Preece and Levy, 2018; Amiruddin, 2022; Bergen et al., 2022). Reading enjoyment has been consistently found to have a significant positive association with students' reading performance and achievement in diverse populations and age groups (Clark, 2011; Shanahan and Lonigan, 2015; Ho and Lau, 2018). Concerning the association between engagement and enjoyment, in a meta-analysis of 52 studies, Guthrie and Wigfield (2000) reported a correlation of $r = 0.74$ which is large by any standards (e.g., Cohen, 1992). Additionally, fostering students' enjoyment of reading is imperative to support continued reading engagement (Merga and Roni, 2018). It has been highlighted that poor attitudes towards reading can lead to disengagement from reading activities (Martínez et al., 2008; Amiruddin, 2022). Furthermore, students' self-perception of reading ability and enjoyment of reading have been identified as strong correlates of reading achievement (Smith et al., 2012). The correlation between reading enjoyment and reading skills represents a reciprocal association, indicating that literacy skills fuel literacy enjoyment, and vice versa (Bergen et al., 2022). It is believed that this association is through increased engagement with reading activities (Hidi et al., 2006). For example, Allgood et al. (2012) conducted an experimental study to increase student engagement with reading and the outcome was significant gains in reading achievement. Shared reading in school has been associated with increasing student learning, engagement, motivation, and enjoyment (Merga, 2017). Moreover, when students are absorbed in the world of the book, they tend to be particularly engaged in their reading activity (Mol and Jolles, 2014). For example, Guthrie and Wigfield (2000) discovered that students who display elevated levels of engagement, specifically in terms of having the ability to choose and exercise autonomy in their reading activities, exhibit a stronger sense of enjoyment towards reading. This pleasurable experience, in turn, facilitates a positive feedback loop, strengthening the level of involvement (Mol and Bus, 2011).

1.1 Reading engagement and disruptive classroom behavior

Student's disruptive behavior can make the classroom atmosphere unsuitable for learning (Stage and Quiroz, 1997). It may result in less time spent on teaching and divert the attention of other pupils, which might have an indirect negative impact on the class's reading engagement and success (Sullivan et al., 2014). Research has demonstrated that the use of effective classroom management strategies is associated with a reduction in disruptive conduct and a corresponding improvement in academic performance (Bradshaw et al., 2010). Examples of good practices are peer-assisted learning Sinclair et al. (2019), the good behavior game (Smith et al., 2012), and the systematic analysis of behavior (Shumate and Wills, 2010). Furthermore, previous research has demonstrated that disruptive behaviors play a crucial role in moderating the relationship between class assignment and reading proficiency in kindergarten, pointing to the potentially detrimental effect on reading acquisition (Coventry et al., 2009). Even more important is the fact that disruptive behaviors have a substantial role in moderating the relationship between students' reading proficiency, their conduct within the educational setting, and the instructional competencies of teachers (Brokamp et al., 2018). Research has demonstrated that the presence of disruptive behavior can lead to adverse psychological effects such as stress, anxiety, annoyance, and even rage on the part of teachers. Poor teacher-student relationships can result in lower expectations and lower-quality instruction, which can affect students' reading success (Hughes et al., 2008). These negative emotions can hinder effective communication and collaboration among individuals, ultimately leading to a possible decline in the quality of education and services provided (Rosenstein and O'Daniel, 2008) and similarly significant decrements in students' reading achievement (Pisecco et al., 2001). Interestingly, the potentially moderating role of a disruptive classroom environment on student engagement with reading activities has only been investigated using linear analytical means assuming an analogous effect across all levels of disruptive behavior. The present study hypothesizes that the relationship between disruptive student behaviors and engagement in reading activities is most likely non-linear and best described by the cusp catastrophe (Cobb and Zacks, 1985). Below there is an analytical account of this thesis.

1.2 Nonlinear dynamics and the cusp catastrophe model

The cusp catastrophe model, a fundamental idea within the field of nonlinear dynamical systems theory, was formulated by René Thom in the 1970s and subsequently popularized by Eric Zeeman. This model plays a crucial role in explaining abrupt and profound shifts in behavior or occurrences, which linear models encounter difficulty in accurately forecasting. The extensive utilization of the cusp catastrophe model across many fields highlights its adaptability and a broad range of applications. The economic model elucidates non-linear associations between predictors and outcomes, encompassing the dynamics of financial markets during times of crisis and the anticipation of pivotal junctures within economic systems (Chen et al., 2014, 2020). Within the field of engineering, the utilization of stability analysis is prevalent in the examination of nonlinear material structures and the anticipation of catastrophic failures resulting from stress-induced conditions (Wang et al., 2011). Furthermore, the utilization of the model has been observed in the domains of public health, and behavioral research, as well as in the comprehension of intricate phenomena such as the dynamics of rangeland ecosystems and fetal heart rate decelerations (Lockwood and Lockwood, 1993; Kikuchi et al., 2006). More recently, several studies in education, educational psychology, and mainstream psychology have employed the cusp catastrophe model. These studies attempted to explain the roles and functioning of motivation (Stamovlasis and Gonida, 2018), problem solving (Stamovlasis and Tsaparlis, 2012), health (Clair, 1998) or public health concerns (Ding-Geng and Chen, 2017) to mention a few.

The functioning of the cusp model is based on the integration of two control parameters, which serve as external factors affecting the system, along with a behavior variable that signifies the current state of the system. As the aforementioned parameters exhibit variability, the system experiences a significant metamorphosis, distinguished by an abrupt transition from one state to another. The sudden transition, referred to as a cusp., takes place along a distinct curve inside the parameter space, highlighting the significant influence of these external inputs in initiating the system's metamorphosis. The model is represented as a three-dimensional surface, frequently exhibiting a

cusp-like form, wherein smooth variations in the control parameters can result in sudden and discontinuous alterations in the behavior variable, a phenomenon referred to as 'bifurcation'. In equation form (Lockwood and Lockwood, 1993), the cusp catastrophe model is described by a potential function $V(y, a, b)$ as follows:

$$V(y\alpha\beta) = \alpha y + \frac{1}{2}\beta y^2 - \frac{1}{4}y^4 \quad (1)$$

With the potential function 'V', state variable 'y', and the asymmetry and bifurcation parameters 'a' and 'b'. The values of the parameters 'a' and 'b', which are considered to move slowly in comparison to y, define the state of the system. As the two control parameters change the behavior evolves either gradually or suddenly depending on when the bifurcation term value enters the so-called critical point for which abrupt and sudden changes in the outcome variable in any direction are expected.

The purpose of the present study was to explore the potentially complex relationship between disordered behavior at school and students' engagement with reading activities given that they enjoy reading. Of particular interest is the role of disordered behavior which we believe moderated the relationship between liking reading and reading engagement. It is hypothesized that its role is moderating but also in a non-linear fashion. That is, moderators are evaluated at different levels within their linear scaling. For disordered and disruptive behavior in the class, this relationship is likely non-linear as reading engagement likely drops to extremely low levels when disruption levels exceed any manageable by the teacher level. Consequently, the relationship between student-disordered behavior in the class and engagement with reading activities is likely better modeled within the cusp catastrophe model for which engagement may likely present itself with abrupt and discontinuous alterations.

2 Method

2.1 Participants

Participants were 2,420 fourth graders who participated in the 2021 PIRLS study from Saudi Arabia. Students were selected using stratified random sampling from 117 schools in the Kingdom. Only Saudi students and those who had complete data participated, thus, listwise deletion was employed. Exclusionary criteria involved international students or students whose native language was not Arabic and those whose achievement was too low to be estimated to avoid floor effects in achievement. There were 1,434 girls (59.3%) and 986 boys (40.7%). Data, methodology, scales, and reports from PIRLS 2021 may be accessed directly at: https://pirls2021.org/.

2.2 Measures

All scales were completed by students. Estimation of internal consistency reliability involved Cronbach's alpha.

2.2.1 Disorderly behavior during lessons

This scale is comprised of five items evaluating the frequency with which disorderly conduct is present in the classroom and is based on student reports. Example behaviors were "students do not listen to what the teacher says" or "there is too much noise for students to work well." (see Supplementary Appendix A). Items were scored using a 4-point rating scale system anchored between the options "never" and "every or almost every lesson." The scale was utilized using its original scoring system which was based on the fit of the Rasch model. The direction of scoring was so that lower scores are indicative of aberrant behavioral patterns. Alpha internal consistency reliability was 0.83.

2.2.2 Students like reading

This scale also completed by students was comprised of 8 items utilizing a r-point scaling system denoting agreement to disagreement. Item content related to the joy of reading, the challenge and learning from reading, etc. (see Supplementary Appendix A). The scale scores using the Rasch model were utilized as per the developer's suggestions. Higher scores were indicative of higher interest and joy from being engaged in reading activities. Alpha internal consistency reliability was 0.81.

2.2.3 Students engaged in reading lessons

This scale included nine items using a 4-point agreement-disagreement scaling system. Sample items were "My teacher gives me interesting things to read," and "My teacher encourages me to say what I think about what I have read" (see Supplementary Appendix A). Higher scores were indicative of higher engagement with reading tasks. The alpha internal consistency reliability of the scale was 0.83.

2.3 Data analyzes

2.3.1 Cusp catastrophe model and prerequisite assumptions

The main assumption of the cusp model is the presence of bimodality or multimodality in the dependent variable suggesting different states of behavior as a function of the asymmetry and bifurcation variables. For this reason, I employed the multimode package (Ameijeiras-Alonso et al., 2021) in R which acts as a toolbox for assessing multimodality by engaging the diptest package (Maechler, 2016) for applying the Hartigan and Hartigan (1985) procedure, and the modeest package (Poncet, 2019) to assess the true number of modes.

The cusp catastrophe model was evaluated using the cusp package in R (Grasman et al., 2009) and variables were standardized as theta scores from the Rasch model were used. Figure 1 displays the main theses of the cusp catastrophe in the context of students' engagement in reading. When levels in the asymmetry variable (namely liking of reading) and the bifurcation variable (disordered behavior in the classroom) are low, the relationship between student reading engagement and disruptive behavior is likely linear and positive as shown in Pattern A. However, when levels of disruption exceed a critical high level, termed the cusp point (point B in the figure), from which the classroom environment is no longer conducive to learning, the cusp model expects that reading engagement becomes unpredictable and is no longer explained using linear terms (see Pattern B). This qualitative description of the reading engagement process provided by its three-dimensional model renders it a potent tool for understanding complex and

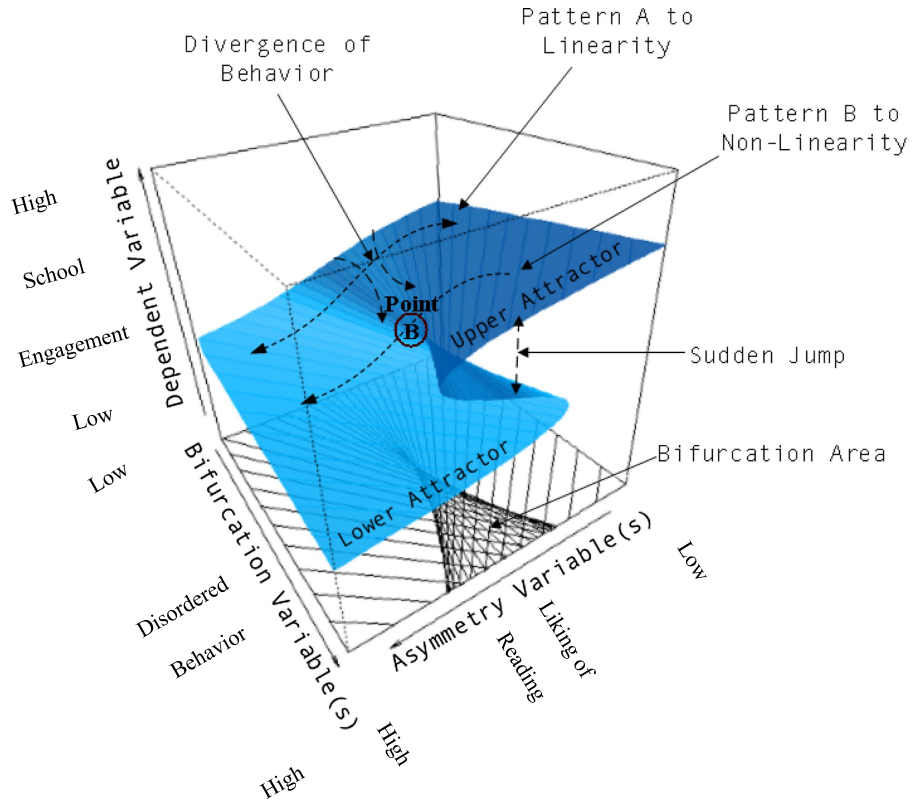


FIGURE 1
Description of the cusp model within the context of student's engagement with reading activities (outcome variable) predicted by the linear effects of students' liking of reading (asymmetry variable) and the splitting effects from having a disordered classroom environment (bifurcation variable).

multivariate educational phenomena (Chen and Chen, 2015). Omnibus model fit was evaluated by contrasting the cusp model with the linear model (as in multiple regression analysis with all predictors entered in one step), and the logistic model (evaluating the behavior of the outcome variable using an S-shaped curve). In particular, the logistic model provides for a competing alternative to the cusp model as it also models nonlinear trajectories. The level of significance was set to 0.01 to account for the relatively large sample size and the correspondingly large amounts of statistical power.

3 Results

3.1 Prerequisite statistical analyzes

Figure 2 displays the findings from the tests of bimodality and multimodality. As shown in Figure 2, upper panel, four modes were identified. First, the conclusion of multimodality was confirmed using Hartigan's dip test for unimodality ($D = 0.084$, $p < 0.001$). As a second step, Silverman's (1981) critical bandwidth test evaluated alternative hypotheses for the presence of more than one mode. All tests up to 3 modes pointed to accepting the alternative hypothesis that a different number of modes was evident. Only when 4 modes was the reference value, the null hypothesis was supported in that the actual number of modalities was not different from four (Critical bandwidth = 0.363, $p = 0.058$). Figure 2, lower panel, displays the sizer plot with the transition between the colors blue and purple indicating a change in the trajectory of behavior from a negative trend to a zero trend, and colors transitioning from purple to red, changes in behavior from a zero trend to a positive trend (Chaudhuri and Marron, 1999). All this information adds evidence to the conclusion of multimodality in the dependent variable.

3.2 Prediction of reading engagement from reading enjoyment and a disordered classroom environment

Table 1 displays global fit statistics from contrasting linear, logistic, and cusp models. As shown in the table, all information criteria values were saliently smaller in the cusp model compared to the linear and logistic comparison models. Further evidence was provided by contrasting the linear and cusp model using a chi-square difference test, which was significant in favor of the latter [χ^2 (3) = 4,698, $p < 0.001$]. Thus, model fit significantly favored the cusp model over competing models.

Table 2 displays the parameters of the cusp model, with all being significantly different from zero. Focusing on the slope terms of the asymmetry and bifurcation variables, the liking of reading was a significant positive predictor of student engagement with reading as expected ($b = 0.331$, $p < 0.001$). Similarly, student disordering in the classroom had a positive slope which is associated with the presence of sudden and unpredictable changes in reading engagement as per

the cusp model premises ($b = 0.112$, $p < 0.001$). Thus, as the asymmetry factor increases, that is, the liking of reading and disordered behavior is at low levels student engagement with reading grows linearly. However, when classroom-disordered behavior grows beyond some critical adaptive point, student engagement with reading takes on various values and becomes unpredictable.

Figure 3, right panel, displays distributions of students' responses at various areas of the lower response surface. As posited by the main theses of the model (e.g., Cobb and Zacks, 1985), bimodality and multimodality are evident at various areas within the response surface with a small number of observations (i.e., $n = 6$) being present within the bifurcation area. The upper left panel of Figure 3 displays the observations as they oscillate from the upper to the lower surface. Observations with "darker" colors are closer to the upper surface and the opposite is true of observations with lighter colors. The larger dots are indicative of coordinates with data from more than one participant. The lower right part of the figure displays the observations as they move from the upper to the lower surface. Last, Figure 4 displays residual versus fitted values for which a slight negative trend is to be expected as was the case with simulated data (Grasman et al., 2009). Collectively all the information corroborates with the idea that the present data were a good fit for the cusp catastrophe model.

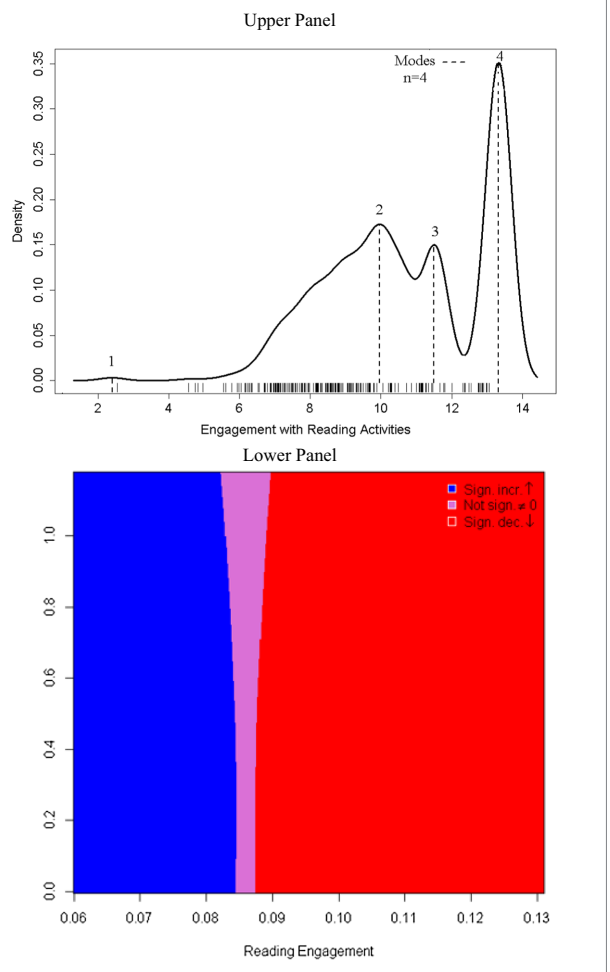


FIGURE 2
Multimodality in student engagement with reading activities at school using density plot (upper panel) and the sizer map (lower panel).

4 Discussion

The purpose of the present study was to explore the potentially complex relationship between disordered behavior at school and students' engagement with reading activities given that they like and enjoy reading. Of particular interest is the role of disordered behavior which, as expected, moderated the relationship between students' liking of reading and reading engagement.

With reading enjoyment serving as the asymmetry variable and disordered student behavior as the bifurcation variable, the cusp catastrophe model offered a sophisticated knowledge of how engagement can fluctuate suddenly and unexpectedly due to the classroom environment. The prevailing scenario from the present findings is that engagement remains stable when disruptive behavior in the classroom increases up to some moderate levels that define a critical point, the cusp point. Beyond that point, any minor increase in students' disruptive behavior is likely associated with a significant and sudden drop in students' engagement with reading activities. This finding adds to the scientific literature that has demonstrated the negative propensities of disruptive behavior in the classroom and suggests tha these effects are more pronounced than what was earlier predicted using the linear model (Stage and Quiroz, 1997; Bradshaw et al., 2010). Empirical studies conductedin educational settings have yielded evidence that supports the presence of nonlinear effects in student engagement (Oliver et al., 2011). For example, scientific studies have demonstrated that levels of engagement can vary significantly and are influenced by factors such as students' positive emotions and adaptive coping strategies (Reschly et al., 2008; Guardino and Fullerton, 2010; Cook et al., 2013). The present results highlight the significance of maintaining a well-structured classroom environment and cultivating a favorable mindset towards reading. Even minor adjustments in these aspects can result in notable improvements in student involvement. This method emphasizes how important it is for the classroom environment and each student's attitudes toward reading to play a part in determining engagement patterns. This approach provides educators with a framework for recognizing and resolving the

TABLE 1 Model comparison across linear and cusp models using descriptive information criteria.

Models tested	Loglikelihood	Parameters	AIC	AICc	BIC
1. Linear	−5836.003	4	11680.010	11680.020	11703.725
2. Logistic	−5776.694	5	11563.390	11563.410	11593.038
3. Cusp	−3486.770	7	6987.540	6987.581	7029.049

Npar, Number of estimated parameters; AIC, Akaike criterion; AICc, Corrected AIC with an adjustment for sample size; BIC, Bayesian information criterion.

TABLE 2 Parameter estimates of the cusp model for the prediction of student engagement with reading activities as a function of student liking of reading (asymmetry var.) and disordered student behavior in class (bifurcation var.).

Terms in Cusp Model	Slope	$LCI_{95\%}$	$UCI_{95\%}$	S.E.	*Z*-test	*p*-value
a_0(Intercept)	−3.015	−3.681	−2.348	0.340	−8.859	<0.001***
a_1(Liking of Reading)	0.331	0.276	0.387	0.028	11.74	<0.001***
b_0(Intercept)	−1.453	−1.550	−1.356	0.050	−29.272	<0.001***
b_1(Student Disordered Behavior)	0.112	0.103	0.121	0.005	24.127	<0.001***
w_0(Intercept)	−3.522	−3.639	−3.405	0.060	−59.004	<0.001***
w_1(Student Engagement)	0.364	0.356	0.373	0.004	82.067	<0.001***

The terms a, b, and w refer to asymmetry, bifurcation, and outcome variables' intercept and slope terms, respectively; Intercepts are denoted with the '0' subscript and slopes with "1".
$LCI_{95\%}$ = Lower 95% Confidence Interval; $UCI_{95\%}$ = Lower 95% Confidence Interval.
***$p < 0.001$, **$p < 0.01$, *$p < 0.05$ for two-tailed tests. The observed *p*-values were further corrected for experimenter-wise error and were still all significant at $p < 0.001$.

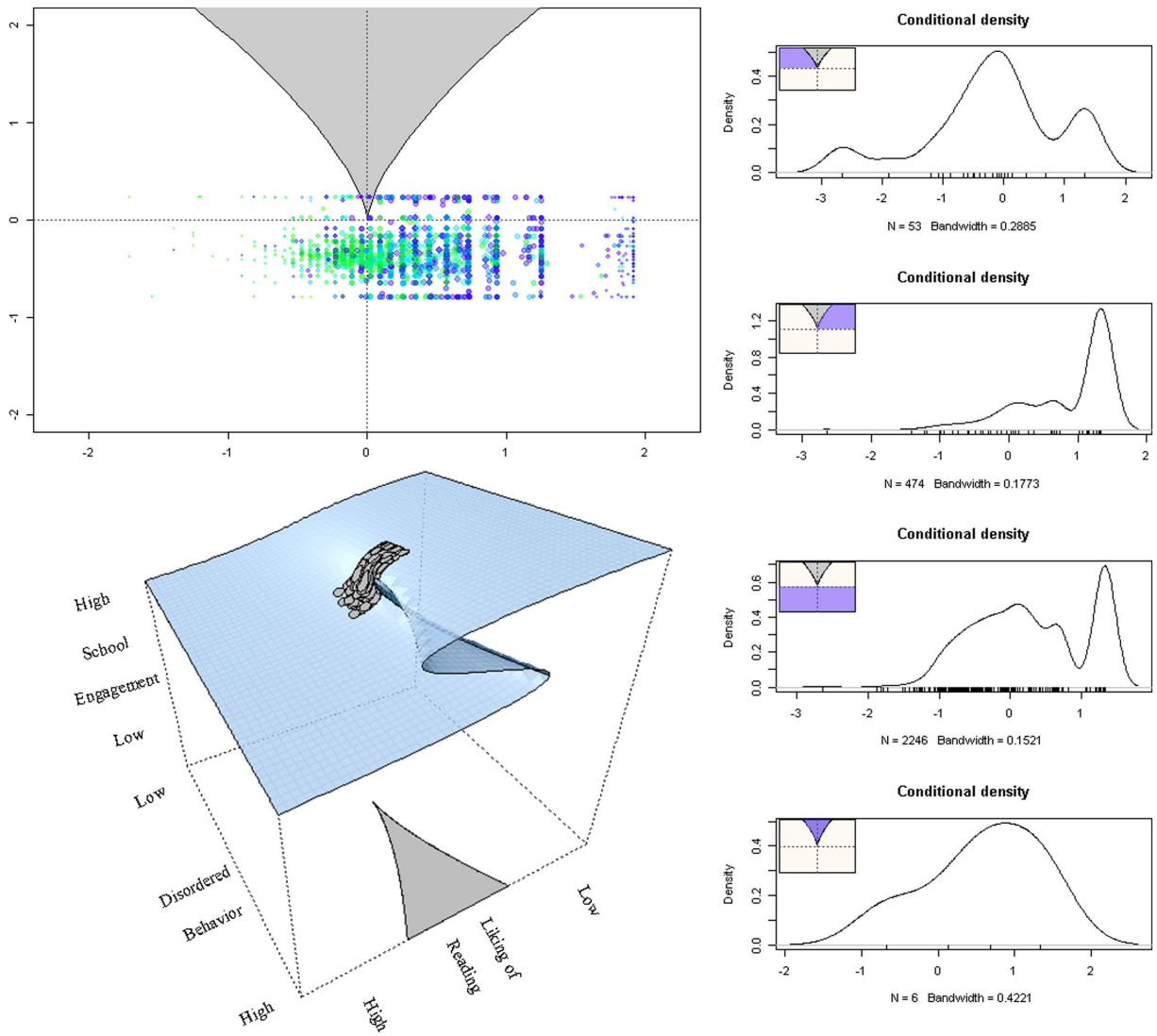


FIGURE 3
The upper left figure displays the lower surface with observations oscillating from the upper to the lower surface. The lower left figure shows observations transitioning across surfaces. The figures to the right show densities at various locations on the response surface. The terms "High" and "Low" refer to levels of the outcome, asymmetry, and bifurcation variables.

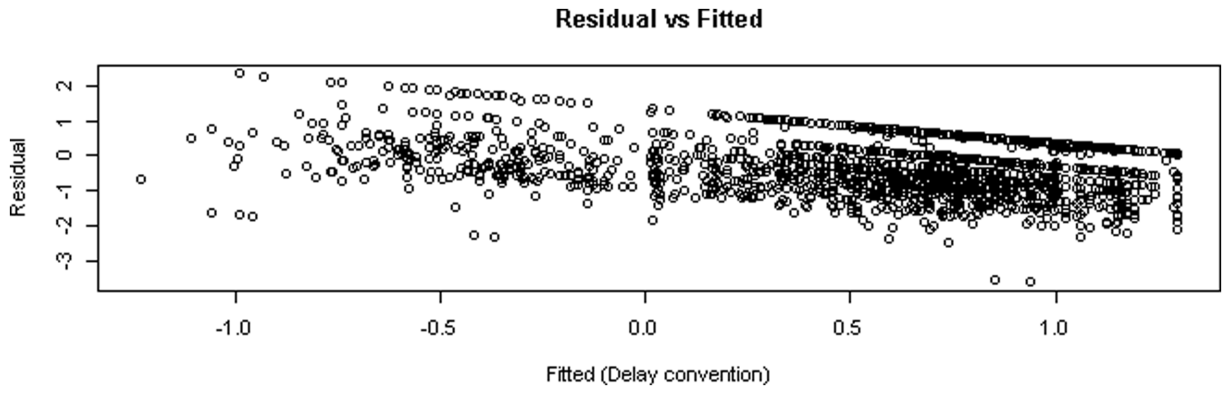


FIGURE 4
Scatterplot of fitted vs. residual values in the cusp catastrophe model.

critical elements that might have an abrupt effect on students' participation in academic activities.

4.1 Study limitations and future directions

Several items related to the cusp catastrophe model contribute to its limitations. First, causality cannot be inferred as a correlational design was utilized and the data represent a snapshot of what was in place during 2021 in schools in the Saudi Arabia Kingdom. Second, the cusp model has been criticized for lacking generalizability as individual and contextual factors vary by classroom and school, thus, the generality of the present findings should be viewed with caution (see Cobb and Zacks, 1985). Third, overfitting the model is a potential risk as simpler analytical models may also fit the specific model and be preferred using the principle of parsimony (Stewart, 1981). Fourth, the self-reported nature of the data are associated with some degree of correlation due to the common method. Thus, the observed relationships may be likely inflated due to medium. Last, the analytical methodology reflected a selection of Cobb's model among other alternatives such as Guastello's polynomial regression. In the future, it will be important to replicate the present findings and extend them by including person-relevant attributes that may act as a buffer against the negative effects of a disruptive classroom environment. Gender differences and social-contextual factors such as student SES and private or public schooling may be important moderators towards understanding such complex educational phenomena.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://pirls2021.org/.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

GA: Conceptualization, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author declares financial support was received for the research, authorship, and/or publication of this article. The author would like to acknowledge Deanship of Scientific Research, Taif university for funding this work.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1346232/full#supplementary-material

References

Afflerbach, P., Cho, B., Kim, J., Crassas, M., and Doyle, B. (2013). Reading: what else matters besides strategies and skills? *Read. Teach.* 66, 440–448. doi: 10.1002/TRTR.1146

Afflerbach, P., Pearson, P. D., and Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *Read. Teach.* 61, 364–373. doi: 10.1598/RT.61.5.1

Allgood, J., Welch, M. B., and Wenner, J. (2012). An examination of the effects of a reading intervention that incorporates student engagement. *Elem. Sch. J.* 113, 57–82.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodríguez-Casal, A. (2021). Multimode: Mode testing and exploring. R package version 1.5. Available at: https://CRAN.R-project.org/package=multimode

Amiruddin, A. (2022). The influence of sq3r technique and students' reading interest towards students' reading comprehension achievement. *J. Soc. Work Educ.* 3, 60–66. doi: 10.52690/jswse.v3i1.273

Baker, L., Dreher, M. J., and Guthrie, J. T. (Eds.). (2000). *Engaging young readers: Promoting achievement and motivation*. Guilford Press. New York, NY

Bergen, E., Hart, S., Latvala, A., Vuoksimaa, E., Tolvanen, A., and Torppa, M. (2022). Literacy skills seem to fuel literacy enjoyment, rather than vice versa. *Dev. Sci.* 26:e13325. doi: 10.1111/desc.13325

Bradshaw, C. P., Mitchell, M. M., and Leaf, P. J. (2010). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes. *J. Posit. Behav. Interv.* 12, 133–148. doi: 10.1177/1098300709334798

Brokamp, S., Houtveen, A., and Grift, W. (2018). The relationship among students' reading performance, their classroom behavior, and teacher skills. *J. Educ. Res.* 112, 1–11. doi: 10.1080/00220671.2017.1411878

Chaudhuri, P., and Marron, J. S. (1999). SiZer for exploration of structures in curves. *J. Am. Stat. Assoc.* 94, 807–823. doi: 10.1080/01621459.1999.10474186

Chen, D. G., Lin, F., Chen, X., and Tang, W. (2014). Cusp catastrophe model: A nonlinear model for health outcomes in nursing research. *Nurs. Res.* 63, 211–220.

Chen, X., and Chen, D. (2015). "Cusp catastrophe modeling in medical and health research," in *Innovative statistical methods for public health data*. eds. D.-G. Chen and J. Wilson (Cham, Switzerland: Springer).

Chen, X., Wang, K., and Chen, D. G. (2020). "Cusp catastrophe regression analysis of testosterone in bifurcating the age-related changes in PSA, a biomarker for prostate cancer," in *Statistical methods for global health and epidemiology: Principles, methods and applications*. eds. X. Chen and D. G. Chen (Cham, Switzerland: Springer), 353–372.

Clair, S. (1998). A cusp catastrophe model for adolescent alcohol use: an empirical test. *Nonlinear Dyn. Psychol.* 2, 217–241. doi: 10.1023/A:1022376002167

Clark, C. M. (2011). Reading enjoyment and motivation: a complex relationship of literacy and socio-emotional factors. *Read. Res. Q.* 46, 315–339.

Cobb, L., and Zacks, S. (1985). Applications of catastrophe theory for statistical modeling in the biosciences. *J. Am. Stat. Assoc.* 80, 793–802. doi: 10.1080/01621459.1985.10478184

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155

Cook, C., Collins, T., Dart, E., Vance, M., McIntosh, K., Grady, E., et al. (2013). Evaluation of the class pass intervention for typically developing students with hypothesized escape-motivated disruptive classroom behavior. *Psychol. Sch.* 51, 107–125. doi: 10.1002/pits.21742

Coventry, W., Byrne, B., Coleman, M., Olson, R., Corley, R., Willcutt, E., et al. (2009). Does classroom separation affect twins' reading ability in the early years of school? *Twin Res. Hum. Genet.* 12, 455–461. doi: 10.1375/twin.12.5.455

Ding-Geng, C., and Chen, X. (2017). Cusp catastrophe regression and its application in public health and behavioral research. *Int. J. Environ. Res. Public Health* 14:1220. doi: 10.3390/ijerph14101220

Duke, N. K., and Pearson, P. D. (2002). "Effective practices for developing reading comprehension" in *What research has to say about reading instruction*. eds. A. E. Farstrup and S. J. Samuels. *3rd* ed (Newark, DE: International Reading Association), 205–242.

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School engagement: potential of the concept, state of the evidence. *Rev. Educ. Res.* 74, 59–109. doi: 10.3102/00346543074001059

Grasman, R. P., van der Maas, H. L., and Wagenmakers, E. J. (2009). Fitting the cusp catastrophe in R: A cusp-package primer. *J. Stat. Softw.* 32, 1–27.

Guardino, C., and Fullerton, E. (2010). Changing behaviors by changing the classroom environment. *Teach. Except. Child.* 42, 8–13. doi: 10.1177/004005991004200601

Guthrie, J. T., and Klauda, S. L. (2008). "Engagement and motivation in reading" in *Handbook of reading research*. eds. A. Farstrup and S. J. Samuels, vol. *4* (Abingdon: Taylor & Francis), 408–432.

Guthrie, J. T., and Wigfield, A. (2000). "Engagement and motivation in reading" in *Handbook of reading research*. eds. M. L. Kamil, P. B. Mosenthal, P. D. Pearson and R. Barr, vol. *III* (Mahwah, NJ: Lawrence Erlbaum Associates), 403–422.

Hartigan, J. A., and Hartigan, P. M. (1985). The dip test of unimodality. *Ann. Stat.* 13, 70–84. doi: 10.1214/aos/1176346577

Hidi, S., Berndt, K., and Pearson, A. D. (2006). Is interest important? A meta-analysis of interest in learning and its effects on student motivation and achievement. *Rev. Educ. Res.* 76, 123–156.

Ho, E., and Lau, K. (2018). Reading engagement and reading literacy performance: effective policy and practices at home and in school. *J. Res. Read.* 41, 657–679. doi: 10.1111/1467-9817.12246

Hughes, J. N., Luo, W., Kwok, O. M., and Loyd, L. K. (2008). Teacher-student support, effortful engagement, and achievement: a 3-year longitudinal study. *J. Educ. Psychol.* 100, 1–14. doi: 10.1037/0022-0663.100.1.1

Kikuchi, A., Unno, N., Horikoshi, T., Kozuma, S., and Taketani, Y. (2006). Catastrophe theory model for decelerations of fetal heart rate. *Gynecol. Obstet. Invest.* 61, 72–9. doi: 10.1159/000088812

Lim, H., Bong, M., and Woo, Y. (2015). Reading attitude as a mediator between contextual factors and reading behavior. *Teach. Coll. Rec.* 117, 1–36. doi: 10.1177/016146811511700116

Lockwood, J., and Lockwood, D. R. (1993). Catastrophe theory: A unified paradigm for rangeland ecosystem dynamics. *J. Range Manag.* 46, 282–288.

Maechler, M. (2016). Diptest: Hartigan's dip test statistic for unimodality – Corrected. R package version 0.75–7. Available at: http://CRAN.R-project.org/package=diptest

Martínez, R., Arıcak, O., and Jewell, J. (2008). Influence of reading attitude on reading achievement: a test of the temporal-interaction model. *Psychol. Sch.* 45, 1010–1023. doi: 10.1002/pits.20348

Merga, M. (2017). Interactive reading opportunities beyond the early years: what educators need to consider. *Aust. J. Educ.* 61, 328–343. doi: 10.1177/0004944117727749

Merga, M., and Roni, S. (2018). Empowering parents to encourage children to read beyond the early years. *Read. Teach.* 72, 213–221. doi: 10.1002/trtr.1703

Mol, S. E., and Bus, A. G. (2011). To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. *Psychol. Bull.* 137, 267–296. doi: 10.1037/a0021890

Mol, S., and Jolles, J. (2014). Reading enjoyment amongst non-leisure readers can affect achievement in secondary school. *Front. Psychol.* 5:1214. doi: 10.3389/fpsyg.2014.01214

Oliver, R., Wehby, J., and Reschly, D. (2011). Teacher classroom management practices: effects on disruptive or aggressive student behavior. *Campbell Syst. Rev.* 7, 1–55. doi: 10.4073/csr.2011.4

Pisecco, S., Wristers, K., Swank, P., Silva, P., and Baker, D. (2001). The effect of academic self-concept on adhd and antisocial behaviors in early adolescence. *J. Learn. Disabil.* 34, 450–461. doi: 10.1177/002221940103400506

Poncet, P. (2019). Modeest: mode estimation. R package version 2.4.0. Available at: https://CRAN R-project.org/package=modeest

Preece, J., and Levy, R. (2018). Understanding the barriers and motivations to shared reading with young children: the role of enjoyment and feedback. *J. Early Child. Lit.* 20, 631–654. doi: 10.1177/1468798418779216

Pressley, M., and Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.

Reschly, A. L., Huebner, E. S., Appleton, J. J., and Antaramian, S. (2008). Engagement as flourishing: the contribution of positive emotions and coping to adolescents' engagement at school and with learning. *Psychol. Sch.* 45, 419–431. doi: 10.1002/pits.20306

Rosenstein, A., and O'Daniel, M. (2008). Invited article: managing disruptive physician behavior: impact on staff relationships and patient care. *Neurology* 70, 1564–1570. doi: 10.1212/01.wnl.0000310641.26223.82

Shanahan, T., and Lonigan, C. J. (2015). *Effective reading instruction for young children: Kindergarten through 3rd grade* (*2nd*). Guilford Press. New York, NY

Shumate, E., and Wills, H. (2010). Classroom-based functional analysis and intervention for disruptive and off-task behaviors. *Educ Treat Child* 33, 23–48. doi: 10.1353/etc.0.0088

Silverman, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. R. Stat. Soc. Series B* 43, 97–99.

Sinclair, A., Gesel, S., and Lemons, C. (2019). The effects of peer-assisted learning on disruptive behavior and academic engagement. *J. Posit. Behav. Interv.* 21, 238–248. doi: 10.1177/1098300719851227

Smith, J., Smith, L., Gilmore, A., and Jameson, M. (2012). Students' self-perception of reading ability, enjoyment of reading and reading achievement. *Learn. Individ. Differ.* 22, 202–206. doi: 10.1016/j.lindif.2011.04.010

Stage, S. A., and Quiroz, D. R. (1997). A meta-analysis of interventions to decrease disruptive classroom behavior in public education settings. *Sch. Psychol. Rev.* 26, 333–368. doi: 10.1080/02796015.1997.12085871

Stamovlasis, D., and Gonida, E. (2018). Dynamic effects of performance-avoidance goal orientation on student achievement in language and mathematics. *Nonlinear Dynamics Psychol. Life Sci.* 22, 335–358.

Stamovlasis, D., and Tsaparlis, G. (2012). Applying catastrophe theory to an information-processing model of problem solving in science education. *Sci. Educ.* 96, 392–410. doi: 10.1002/sce.21002

Stewart, I. (1981). Catastrophe theory in psychology: a user's guide. *Int. J. Man-Mach. Stud.* 14, 385–399.

Sullivan, A. L., Van Norman, E. R., and Klingbeil, D. A. (2014). Exclusionary discipline of students with disabilities: student behaviors and school characteristics. *J. Emot. Behav. Disord.* 22, 212–222.

Unrau, N., and Schlackman, J. (2006). Motivation and its relationship with reading achievement in an urban middle school. *J. Educ. Res.* 100, 81–101. doi: 10.3200/JOER.100.2.81-101

Wang, T. T., Yan, X. Z., Yang, H. L., and Yang, X. J. (2011). Stability analysis of the pillars between bedded salt cavern gas storages by cusp catastrophe model. *Sci. China Technol. Sci.* 54, 1616–1623.

Wigfield, A., and Guthrie, J. T. (1997). Relations of children's motivation for reading to the amount and breadth of their reading. *J. Educ. Psychol.* 89, 420–432. doi: 10.1037/0022-0663.89.3.420

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 23 January 2024
DOI 10.3389/fpsyg.2023.1283686

OPEN ACCESS

EDITED BY
Nikolaos Tsigilis,
Aristotle University of Thessaloniki, Greece

REVIEWED BY
Tong Wu,
Riverside Insights™, United States
Hanan Ghamdi,
Education and Training Evaluation Commission (ETEC), Saudi Arabia

*CORRESPONDENCE
Faye Antoniou
✉ fayeantoniou@gmail.com

RECEIVED 26 August 2023
ACCEPTED 18 December 2023
PUBLISHED 23 January 2024

CITATION
Antoniou F and Alghamdi MH (2024) Principal goals at school: evaluating construct validity and response scaling format.
Front. Psychol. 14:1283686.
doi: 10.3389/fpsyg.2023.1283686

Principal goals at school: evaluating construct validity and response scaling format

Faye Antoniou[1]* and Mohammed H. Alghamdi[2]

[1]Department of Educational Studies, National and Kapodistrian University of Athens, Athens, Greece, [2]Department of Self-Development Skills, King Saud University, Riyadh, Saudi Arabia

The purpose of the present study was to test the efficacy and appropriateness of the 4-point response option of the Principal's Goals Scale of the SASS (1999–2000) survey. Competing dichotomous models with various conceptualizations were constructed and tested against the original polytomous conceptualization. Participants were 8,524 principals from whom 64% were males and 36% females. Principals' goals were assessed using a 6-item scale anchored across points reflecting proximity to achieving a goal. The original polytomous conceptualization was contrasted to a dichotomous two-pole conceptualization using a model with freely estimated discriminations (two-parameter logistic model, 2PL) as well as the Rasch model assuming equal discrimination parameters. Results indicated that the 2PL dichotomous model provided the most optimal model fit. Furthermore, item-related, and person-related estimates pointed to enhanced accuracy and validity for the dichotomous model conceptualization compared to the polytomous model. It is suggested that a dichotomous scaling system is considered in subsequent measurements of the scale as a means of enhancing the accuracy and validity of the measured trait.

KEYWORDS

principal goals, SASS survey, response scaling, item response theory, collapsing rating scale categories

1 Introduction

Principals of schools should make it a priority to develop ambitious objectives for their institutions since doing so may have a beneficial effect on many facets of the learning environment and the results for students. Creating a culture in which high expectations are the norm among instructors, students, and parents may be accomplished by setting lofty objectives. According to research conducted by Jussim and Harber (2005), having high expectations from instructors has a beneficial effect on the academic performance of their students. When administrators set lofty objectives for their schools, they inspire every member of the school community to strive for greatness, which in turn leads to an increase in both effort and engagement. In addition, research has indicated that schools with strong leadership and clear objectives tend to have greater levels of student accomplishment (Hallinger and Heck, 1996; Leithwood et al., 2004). There is a correlation between principals who have high standards for academic success and those who provide an atmosphere of support for both teachers and students (Bozkurt, 2023; Perkasa et al., 2023). This correlation contributes to enhanced learning outcomes.

To boost teacher retention rates and promote professional development programs, setting ambitious targets may be quite helpful. The development of a feeling of professional progress

and happiness in one's work is facilitated when administrators articulate an aspiration for academic superiority and provide teachers with the resources necessary to realize that aspiration. According to Hanushek et al. (2004), this, in turn, serves to contribute to the overall quality of education and helps the school retain excellent educators inside the institution. Furthermore, ambitious objectives inspire principals and their teams to seek out creative techniques and apply evidence-based solutions. According to Leithwood et al. (2008), principals can create positive changes in teaching techniques, curriculum design, and school-wide. Policy by cultivating a culture of continuous improvement in their schools. This ultimately results in improved educational experiences for students. According to several studies, one of the ways in which administrators may realize their lofty objectives is by incorporating the stakeholders in the process, as well as the school community, the parents, and the organizations in the surrounding area. According to Epstein (2001), stakeholders, in particular, have the potential to provide resources, opportunities, and enriched experiences, as well as further help in the development of a productive and collaborative atmosphere that contributes to students' overall well-being and academic performance.

At present, several national and international studies have investigated the role and functioning of principals as well as the consequences of their actions. One such study is the "School and Staffing Survey" which mainly collects information from principals regarding school functioning, their roles, and responsibilities as well as their perceived obstacles and barriers to achieving their goals. In the present study, we focus on the principal's goals as we target at re-examining the psychometric qualities of the specific instrument. Besides reliability and construct validity, we are additionally interested in the response scaling system employed as it deviates markedly from Likert-type or frequentist systems. Thus, what is least known, is the efficacy of the response scaling format as the current 4-point scaling system could be suboptimal compared to other available systems, e.g., a dichotomous conceptualization. Currently, scoring includes summed responses of the original 4-point scaling system and validity studies have utilized the total score as a means of estimating total scores. If, however, the current scaling response option proves to be suboptimal, then the associated total scores will have to be revised accordingly in subsequent international measurements.

The literature on survey methods (e.g., Tourangeau et al., 2000) suggests that there are at least three salient contributing factors to consider revising a scale system namely, alignment with other measures, infrequent use of some rating scale options, and conceptual redundancy (Rutkowski et al., 2019). The first refers to harmonizing the scale's definition with those of other instruments that are valid or are considered gold standards. Harmonizing answer categories becomes important when researchers want to compare their findings to those of prior studies or make links between other dimensions (Dusen and Nissen, 2020). Researchers may compare and more easily integrate their results by ensuring uniformity and compatibility across studies by compressing answer possibilities. The second reason for collapsing categories refers to when certain choices are infrequently used (Groves et al., 2011). In many situations, collapsing facilitates data analysis by minimizing the number of categories and enhancing interpretability and statistical power. Response choices that are rarely chosen may not provide useful data or may impede analyses by leaving blank cells or sparse categories (Krosnick and Fabrigar, 1997; Agresti, 2013) as is the case with the omnibus chi-square test that evaluates global model fit. The third refers to the phenomenon when adjacent categories are conceptually similar to the extent that their differentiation is neither clearly defined nor easily attained, thus threatening the reliability of measurement (Embretson and Reise, 2000). On the other hand, the disadvantages of collapsing categories in a rating scale have been a reduction of power (Strömberg, 1996), problems with model convergence (Savalei and Rhemtulla, 2013), distorted model fit (Jeong and Lee, 2016), and loss of reliability and information (Embretson and Reise, 2000; Revilla et al., 2017). Applications of revising scale systems have utilized the constructs of bullying (Rutkowski et al., 2019), personality (Wetzel and Carstensen, 2014), disability status (Dadaş et al., 2020), academic misconduct (Royal et al., 2015), and health status (Williams et al., 2009). The purpose of the present study was to test the efficacy and appropriateness of the 4-point response option of the Principal's Goals Scale of the SASS survey. Competing dichotomous models with various conceptualizations were constructed and tested against the original polytomous conceptualization.

2 Methods

2.1 Participants and procedures

Participants were 8,524 principals who participated in the School and Staffing Survey during 1999–2000. There were 5,481 males (64.3%) and 3,043 females (35.7%). Most principals were above 50 years old (53.7%). There were 348 Hispanic principals representing 4.1% of the total sample. Regarding race, 87.1% were white followed by black (9.9%), American Indian (1.8%), and Asian (1.2%). All but 1.6% had at least a Master's degree. The sampling frame in SASS used the Common Core of Data (CCD) file that includes all elementary and secondary schools in the USA. Sampling in the SASS involved school selection using a probability proportionate to the square root of the number of teachers. Data collection was performed by the U.S. Census Bureau using advance and follow-up letters to the schools and the mode of data collection was computer-assisted telephone interviewing.

2.2 Measure

The principal's goals scale (see Appendix 1) is a six-item scale anchored between a 4-point scaling format ranging from a goal that is far or close to being reached. The potential nominal type scaling with ordered but likely non-equidistant options was a primary motivating factor for evaluating the instrument's response scaling system. Furthermore, scale selection was based on utility as there were 190 published papers or presentations using the specific instrument, with reports confirming adequate levels of reliability and validity (e.g., Blank, 1994).

2.3 Data analyses

2.3.1 Construct validity and person consistency

Data were analyzed using Item Response Theory (IRT) and by employing the Graded Response Model (Samejima, 1969; Muraki, 1992) which is appropriate for polytomous data and a series of models for dichotomous items, namely the Rasch model and the 2-parameter IRT

TABLE 1 Model fit for principal's goals scale using polytomous and dichotomous models.

Model tested	Chi-square	D.F.	value of *p*	RMSEA	AIC	BIC	Omega
M1. Polytomous Graded	26674.39***	4,071	<0.001	0.03	95641.16	95810.37	0.738
M2. Dichotomous-2PL	197.84***	51	<0.001	**0.02**	**47486.52**	**47571.13**	0.652
M3. Dichotomous Rasch	1186.02***	57	<0.001	0.05	48447.57	48489.88	0.453

D.F., Degrees of freedom; RMSEA, Root Mean Squared Error of Approximation; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; Omega, index of internal consistency reliability. Bold values indicate optimal model with the smallest values in the information criteria (AIC, BIC), the RMSEA and the chi-square statistic. ***$p < 0.001$.

model (2PL). Besides the polytomous model, a dichotomy of the 4-scaling system format was created by aggregating the two positive against the two negative responses. Model fit was evaluated using the omnibus chi-square test and the Root Mean Square Error of Approximation (RMSEA). Further local tests included item misfits using chi-square tests and tests of local dependency using the LD index. Given the polytomous nature of the original scaling system, another means of examining scaling appropriateness was the equidistant index (Spratto, 2018), which evaluates the difference between adjacent thresholds, assuming equal distances between rating scale options. Given that thresholds are evaluated in logits, the expected value of the null hypothesis of no differences is equal to zero logits. Although the scaling system deviates markedly from other Likert-type conceptualizations, it was important to examine whether the conceptual distance between "just beginning" and "long way to go," assuming this is the low goal attainment pole, was equivalent to the distance between the "almost there" and "reached our goal" options. Threshold non-equivalence would have implications for psychometrics as the scaling would no longer be considered on the interval scale but should be viewed either as ordered data or even at the nominal level.

Further tests for determining the appropriateness of the scaling system involved the examination of 108 person location fluctuations around the latent trait, termed Person Discriminal Dispersion (PDD) (Ferrando, 2007, 2009; Ferrando and Navarro-González, 2021) which refers to the consistency of the response patterns of individuals about variable item locations (Ferrando, 2016). Well-fitted participants have low values in their discriminal dispersion showing enhanced consistency (Ferrando, 2019). Ferrando and Navarro-González (2020) developed the R package InDisc to provide sample-based estimates of both global fit and person dispersion estimates (R Core Team, 2018). In the present study, we contrasted average estimates of person dispersion between polytomous and dichotomous conceptualizations as a means of evaluating the consistency of the person trait estimates.

2.3.2 Internal consistency reliability

It was assessed using Marginal reliability in light of the recommendations disfavoring Cronbach's alpha as being a low-bound estimate (Sijtsma and Molenaar, 1987; Sijtsma, 2009). Estimates were 0.76 for the polytomous model 0.56 for the 2PL dichotomous conceptualization and 0.45 for the dichotomous conceptualization with fixed slopes (Rasch model).

3 Results

3.1 Model fit as a function of different response scaling formats

A Graded Item Response model was fit to the data as per the original conceptualization. As shown in Table 1, the omnibus chi-square test was significant but unstandardized residuals (i.e., RMSEA) were within the normal range (i.e., 3%). A visual analysis of the items' category curves, however, showed substantial underrepresentation of the "just beginning" category suggesting it was not by itself constructive for measurement purposes (see Figure 1). This finding had significant implications for rating scale equivalence. As shown in Table 2, the conceptual non-equivalence between adjacent thresholds was confirmed as the two poles occupied significantly different spaces across theta. On items 1, 2, 4, and 5, the positive sign of the equidistance index suggests that the distance in thresholds 2 and 3 is significantly larger compared to that of thresholds 1 and 2. Thus, the threshold non-equivalence testing provided some evidence of the lack of optimal functionality of the scaling system.

In light of the above findings on omnibus model fit and threshold non-equivalence, the two adjacent content categories "just beginning" and "long way to go" were aggregated to define the first level of a dichotomy (i.e., zero) with the categories "almost there" and "reached our goal" representing the next category (i.e., one). As shown in Table 1, the smallest chi-square value was reserved by the 2PL model, although the chi-square estimate was significant signaling the expected excessive levels of power. Unstandardized residuals were 2% suggesting "exact model fit" as per MacCallum et al. (1996) recommendations. The second-best model was the dichotomous Rasch model with, however, a significant misfit over the 2PL model by freeing the estimation of 6 discrimination parameters. Given that models were nested, a chi-square difference test pointed to the superiority of the dichotomous 2PL model compared to the Rasch model [$\Delta_{Chi-square}(6) = 988.180$, $p < 0.001$]. In other words, fixing the discrimination parameters to unity was associated with 988 units of model misfit. The polytomous-graded model was by far the worst estimated model. Noteworthy, RMSEA was still acceptable. Thus, global statistical criteria favored a dichotomous response option with two poles as being the most parsimonious and errorless conceptualization for the measurement of principals' attitudes toward their school's goals.

3.2 Item fit statistics: response patterns and residual correlations

Tests of local dependency showed non-significant residual correlations for only the dichotomous 2PL model (see Table 3). Both the dichotomous Rasch model and the polytomous model were associated with significant residual correlations; for the polytomous model, the residual correlations were extended to all pairs of items. For the dichotomous Rasch model, residual correlations were significant across all pairs except items 1 and 3; items 3 and 4; and items 3 and 5. Residual correlations represent a significant obstacle to the validity of person scores as they violate an important prerequisite assumption of the IRT modeling. Furthermore, the fact that two items correlate with

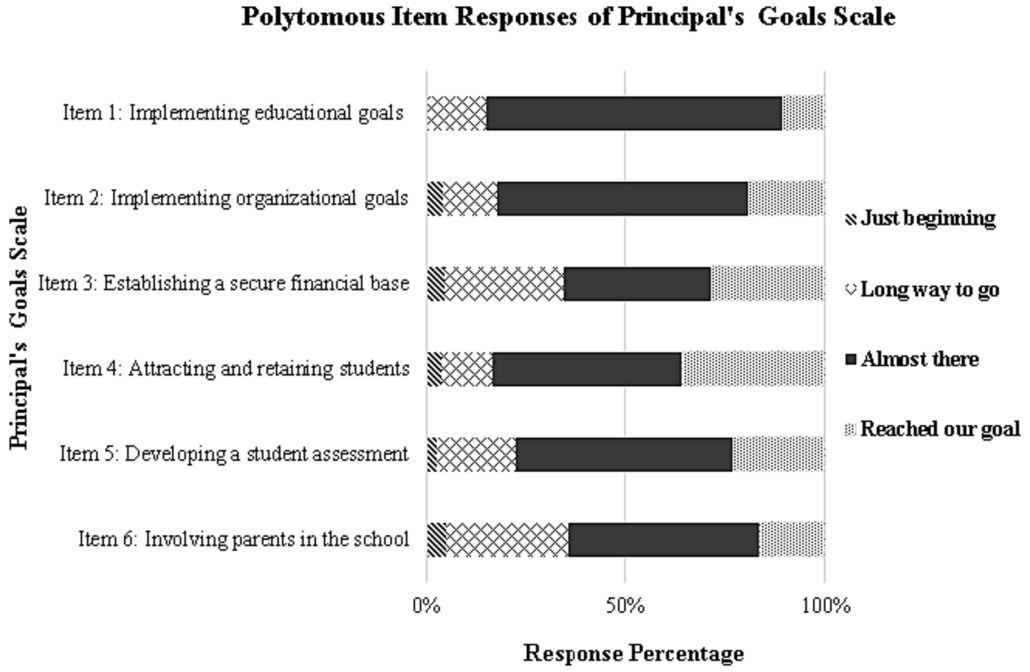


FIGURE 1
Frequency of response categories of the principal's goal scale.

TABLE 2 Equidistance between thresholds in the principal's goals scale.

Item	Discrimination	Threshold/S.E.		Difference/S.E. of Diff		*Z*-test	Equidistance index
Item 1	2.368	−5.270	0.120	–	–	–	–
		−2.639	0.069	1.111	0.034	-	-
		3.726	0.088	2.688	0.042	34.382***	1.577
Item 2	1.821	−4.417	0.086	–	–	–	–
		−2.274	0.051	1.177	0.038	–	–
		2.167	0.049	2.439	0.044	26.379***	1.262
Item 3	0.929	−3.594	0.065	–	–	–	–
		−0.786	0.028	3.023	0.109	–	–
		1.063	0.030	1.990	0.063	12.545***	−1.032
Item 4	1.467	−4.665	0.098	–	–	–	–
		−2.236	0.048	1.656	0.066	–	–
		0.815	0.034	2.080	0.050	6.446***	0.424
Item 5	1.385	−3.780	0.067	–	–	–	–
		−1.504	0.036	1.643	0.051	–	–
		1.652	0.037	2.279	0.049	12.362***	0.635
Item 6	1.327	−3.969	0.071	–	–	–	–
		−0.817	0.031	2.375	0.066	–	–
		2.099	0.041	2.197	0.050	3.06**	−0.178

The equidistance index evaluates the difference between pairs of adjacent thresholds (i.e., 1 and 2 vs. 2 and 3). ***$p < 0.001$; **$p < 0.01$.

each other at the level of variance not explained by the latent construct is both problematic and creates interpretation issues. Thus, collectively, all the evidence pointed to the superiority of the dichotomous 2PL model over the original polytomous model as a more parsimonious and valid assessment of the principal's goals at school.

Further tests of local model fit (i.e., at the item level) utilized the chi-square test to evaluate discrepancies between observed and expected response patterns. As shown in Table 3, the only model for which items fitted the data properly was the dichotomous, 2PL model; both the polytomous and the dichotomous Rasch conceptualization

TABLE 3 Between item residual correlations for principal's goals scale across tested models.

Principal Scale Items	Item 1	Item 2	Item 3	Item 4	Item 5	χ^2/D.F.
Polytomous data-graded model						
Item 1: implementing educational goals	–	–	–	–	–	123.27***/32
Item 2: implementing organizational goals	68.50***	–	–	–	–	99.93***/35
Item 3: establishing a secure financial base	27.40**	35.80***	–	–	–	200.97***/41
Item 4: attracting and retaining students	47.70**	38.20**	68.50***	–		170.85***/39
Item 5: developing a student assessment	34.80***	44.70**	25.40**	24.00**	-	86.38***/37
Item 6: involving parents in the school	36.50**	42.40**	22.00**	38.20***	40.10***	103.32***/38
Dichotomous data-rasch model						
Item 1: implementing educational goals	–	–	–	–	–	356.39***/5
Item 2: implementing organizational goals	483.20***		–	–	–	200.79***/5
Item 3: establishing a secure financial base	0.30	7.40*	–	–	–	45.78***/5
Item 4: attracting and retaining students	138.40***	70.30***	39.80***	–	–	127.04***/5
Item 5: developing a student assessment	178.40***	92.40***	6.20	28.80***	–	48.90***/5
Item 6: involving parents in the school	131.10***	44.60***	1.80	65.00***	40.60***	46.40***/5
Dichotomous data-2PL						
Item 1: implementing educational goals	–	–	–	–	–	19.39**/4
Item 2: implementing organizational goals	1.80		–	–	–	17.77**/4
Item 3: establishing a secure financial base	7.00	0.00	–	–	–	29.11**/4
Item 4: attracting and retaining students	3.20	2.80	37.30***	–	–	26.71**/4
Item 5: developing a student assessment	−0.70	−0.30	2.70	0.80	–	21.83**/4
Item 6: involving parents in the school	−0.10	5.60	−0.60	7.00	1.10	21.21**/4

***$p < 0.001$; **$p < 0.01$; *$p < 0.01$. Non-significant values are indicative of good model fit.

were associated with significantly elevated misfit as evidenced by the very large chi-square values.

3.3 Contrasting rating scale systems using person-based consistency in theta

As described above, estimates of person consistency on latent scores were evaluated in the different scaling systems using the R package InDisc (Ferrando and Navarro-González, 2020). The package provides estimates of global fit such as chi-square statistics and descriptive fit indices such as the Tucker Lewis Index (TLI) and RMSEA. Furthermore, average PDD values are also estimated. When contrasting polytomous versus dichotomous conceptualizations, results indicated a good model fit for only the dichotomous conceptualization [Chi-square(9) = 212.16, TLI = 0.935, RMSEA = 0.056] but not the polytomous one [Chi-square(9) = 442.46, TLI = 0.896, RMSEA = 0.109]. More so, average indices of personal dispersion were 0.51 for the dichotomous conceptualization and 0.66 for the polytomous one. Knowing that lower values are indicative of higher consistency, the dichotomous 2PL model is most likely preferred over the polytomous model.

4 Discussion

The purpose of the present study was to test the efficacy and appropriateness of the 4-point response option of the Principal's Goals Scale of the SASS survey, with the additional goal of proposing the testing rather than implied psychometric properties of instruments and specifically the functioning of their rating scale. Competing dichotomous models with various conceptualizations were constructed and tested against the original polytomous conceptualization.

The present study found that the originally formed scaling system of the principal's goal scale was not associated with an optimal model fit and measurement precision using both scale-level and person-level criteria. Two alternative dichotomous scaling systems were tested, one with free and one with fixed (Rasch) discrimination parameters with the freely estimated 2PL model being associated with the most optimal model fit. The evidence overwhelmingly favored the dichotomous 2PL model as evidenced using fewer numbers of Guttman-related errors using the person dispersion estimates, enhanced amounts of information, and significantly improved model fit. The number of response categories for self-reports of pain interference was investigated in a study by Cook et al. (2010) which found that fewer response categories, as few as five or six, may function as well as the 11 response categories that are conventionally used. However, the results are preliminary since the number of response categories presented was not manipulated in the study design. Therefore, future research should compare the reliability and validity of scores based on the original number of response categories versus a presentation with fewer response options. When scoring assessments, Dusen and Nissen (2020) advised sparingly using data manipulations and keeping all answer categories unless there was a compelling reason to collapse them. They spoke about how to experimentally test for the two probable causes of falling answer categories: loss of utilization and redundancy.

The difficult process of updating a scale system illustrates the scientific community's ongoing efforts to accurately reflect and quantify psychological factors. This project requires balancing dependability and validity. Empirical research relies on reliability and validity, the foundations of robust measurement. Redefining clinical levels by adjusting cutoff values and threshold estimates shows how theory and measurement interact. Modifying scaling systems, however difficult as a task, likely improves construct validity. This improves the accuracy of score-based inferences and conclusions. For example, a depression or anxiety scale may revise its scaling system to redefine the clinical levels of these constructs. The process of revising scaling systems in this particular context functions as a means to recalibrate the operational concepts that form the foundation of these constructions. Its significance cannot be overstated since these latent constructs need to account for the fluctuations of the diagnostic criteria as they are oftentimes altered as they are informed by new empirical findings.

4.1 Study limitations

The present study is limited for several reasons. First, the use of PDDS is rather new and reflects a rather underexplored aspect of person fit. As Ferrando and Navarro-González (2020) stated, the sensitivity of PDD estimates is a function of test length with larger tests having enhanced confidence in the stability of the estimated parameters. Second, models were likely overpowered with $n = 8{,}500$ participants, thus, global fit statistics are likely inflated for these reasons leading to rejections of model fit, even when discrepancies between hypothesized and observed models are not large. The data used pertain to a national database and an instrument that was mostly used between 1999 and 2010, thus, later inferences about the instrument cannot be made. Last, the comparisons between models should take into account the fact that models with fewer categories are artificially inflated for the better as items become more similar, thus, models cannot be contrasted in terms of, e.g., precision as such a finding is attributed to collapsing categories.

4.2 Conclusions and future directions

Before a choice can be taken, the finding that nearby categories need to be merged into a single category has to be verified using other data sets. Researchers must evaluate the final scaling system to ensure that it is relevant, accurately represents the data, and does not compromise the content validity of the study. A collapsing based only on statistical criteria alone is not suggested since low frequency should not be the basis for collapsing; certain significant and useful metrics have a low frequency in the population, but this should not be the main reason for collapsing.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://nces.ed.gov/surveys/sass/question9900.asp.

Ethics statement

The studies involving humans were approved by all this was arranged by NCES at: https://nces.ed.gov/surveys/sass/question9900.asp. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

FA: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. MA: Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Writing – review & editing, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors extend their appreciation to the Deputyship for Research and Innovation, "Ministry of Education" in Saudi Arabia for funding this research (IFKSUOR3-420-1).

Acknowledgments

We would like to acknowledge the support of Sideridis for assisting with the programming of the data analysis steps.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons. Hoboken, New Jersey.

Blank, R. K. (1994). Improving reliability and comparability of NCES data on teachers and other education staff. In schools and staffing survey (SASS). Paper presented at Meetings of the American Statistical Association (NCES 94–01) (pp. 37–50). U.S. Department of Education. Project Officer, Dan Kasprzyk. Washington, DC: NCES Working Paper.

Bozkurt, B. (2023). Social justice leadership as a predictor of school climate. *Pedagog. Res.* 8:em0160. doi: 10.29333/pr/13078

Cook, K. F., Cella, D., Boespflug, E. L., and Amtmann, D. (2010). Is less more? A preliminary investigation of the number of response categories in self-reported pain. *Patient Relat. Outcome Meas.* 2010, 9–18. doi: 10.2147/PROM.S7584

Dadaş, Ö. F., Gökmen, D., and Köse, T. (2020). The effect of different strategies for combining disordered thresholds on Rasch model fit. [Sırasız Eşik Değerlerinin Birleştirilmesinde Farklı Stratejilerin Rasch Modeline Uyum Üzerindeki Etkisi]. *Türkiye Klinikleri Biyoistatistik* 12, 53–69. doi: 10.5336/biostatic.2019-72009

Dusen, M. E., and Nissen, M. E. (2020). Investigating response categories of the Colorado learning attitudes about science survey. *Int. J. Sci. Educ.* 42, 543–557. doi: 10.1080/09500693.2019.1717416

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Epstein, J. L. (2001). *School, family, and community partnerships: Preparing educators and improving schools*. Westview Press, New York.

Ferrando, P. J. (2007). A Pearson-type-VII item response model for assessing person fluctuation. *Psychometrika* 72, 25–41. doi: 10.1007/s11336-004-1170-0

Ferrando, P. J. (2009). A graded response model for measuring person reliability. *Br. J. Math. Stat. Psychol.* 62, 641–662. doi: 10.1348/000711008X377745

Ferrando, P. J. (2016). An IRT modeling approach for assessing item and person discrimination in binary personality responses. *Appl. Psychol. Meas.* 40, 218–232. doi: 10.1177/0146621615622633

Ferrando, P. J. (2019). A comprehensive IRT approach for modeling binary, graded, and continuous responses with error in persons and items. *Appl. Psychol. Meas.* 43, 339–359. doi: 10.1177/0146621618817779

Ferrando, P. J., and Navarro-González, D. (2020). InDisc: an R package for assessing person and item discrimination in typical-response measures. *Appl. Psychol. Meas.* 44, 327–328. doi: 10.1177/0146621620909901

Ferrando, P. J., and Navarro-González, D. (2021). A multidimensional item response theory model for continuous and graded responses with error in persons and items. *Educ. Psychol. Meas.* 81, 1029–1053. doi: 10.1177/0013164421998412

Groves, R. M., Fowler, F. J. Jr., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2011). *Survey methodology.* Hoboken, New Jersey: John Wiley & Sons.

Hallinger, P., and Heck, R. H. (1996). Reassessing the principal's role in school effectiveness: a review of empirical research, 1980–1995. *Educ. Adm. Q.* 32, 5–44. doi: 10.1177/0013161X96032001002

Hanushek, E. A., Kain, J. F., and Rivkin, S. G. (2004). Why public schools lose teachers? *J. Hum. Resour.* 39, 326–354. doi: 10.2307/3559017

Jeong, H., and Lee, W. (2016). The level of collapse we are allowed: comparison of different response scales in safety attitudes questionnaire. *Biom. Biostat. Int. J.* 4, 128–134. doi: 10.15406/bbij.2016.04.00100

Jussim, L., and Harber, K. D. (2005). Teacher expectations and self-fulfilling prophecies: knowns and unknowns, resolved and unresolved controversies. *Personal. Soc. Psychol. Rev.* 9, 131–155. doi: 10.1207/s15327957pspr0902_3

Krosnick, J. A., and Fabrigar, L. R. (1997). Designing rating scales for effective measurement in surveys. In:Lars Lyberg, Paul Biemer, Martin Collins, LeeuwEdith De, Cathryn Dippo and Norbert Schwarzet al, *Survey measurement and process quality* (pp. 141–164). London: John Wiley & Sons.

Leithwood, K., Day, C., Sammons, P., Hopkins, D., and Harris, A. (2008). *Successful school leadership: What it is and how it influences student learning*. The Wallace Foundation. New York.

Leithwood, K., Louis, K. S., Anderson, S., and Wahlstrom, K. (2004). *How leadership influences student learning: A review of research for the learning from leadership project*. London: DfES.

MacCallum, R. C., Browne, M. W., and Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychol. Methods* 1, 130–149. doi: 10.1037/1082-989X.1.2.130

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Appl. Psychol. Meas.* 16, 159–176. doi: 10.1177/014662169201600206

Perkasa, R. P., Suriansyah, A., and Ngadimun, N. (2023). The correlation between the managerial competence of school principal, school climate, and teacher's work spirit with the work commitment at private high school teacher in Banjarmasin. *Int. J. Soc. Sci. Human Res.* 6, 272–280. doi: 10.47191/ijsshr/v6-i1-37

R Core Team. (2018). *R: A language and environment for statistical computing [computer software manual]*. Vienna, Austria: R Core Team.

Revilla, M., Toninelli, D., and Ochoa, C. (2017). An experiment comparing grids and item-by-item formats in web surveys completed through PCs and smartphones. *Tele. Inform.* 34, 30–42.

Royal, K., and Flammer, K. (2015). Measuring academic misconduct: evaluating the construct validity of the exams and assignments scale. *Am. J. Appl. Psychol.* 4, 58–64. doi: 10.11648/j.ajap.s.2015040301.20

Rutkowski, L., Svetina, D., and Liaw, Y. L. (2019). Collapsing categorical variables and measurement invariance. *Struct. Equ. Model. Multidiscip. J.* 26, 790–802. doi: 10.1080/10705511.2018.1547640

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika* 34, 1–97. doi: 10.1007/BF03372160

Savalei, V., and Rhemtulla, M. (2013). The performance of robust test statistics with categorical data. *Br. J. Math. Stat. Psychol.* 66, 201–223. doi: 10.1111/j.2044-8317.2012.02049.x

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* 74, 107–120. doi: 10.1007/s11336-008-9101-0

Sijtsma, K., and Molenaar, W. I. (1987). Reliability of test score in nonparametric item response theory. *Psychometrika* 52, 79–97. doi: 10.1007/BF02293957

Spratto, E. M. (2018). In search of equality: Developing an equal interval Likert response scale. Dissertations 172 Available at: https://commons.lib.jmu.edu/diss201019/172

Strömberg, U. (1996). Collapsing ordered outcome categories: a note of concern. *Am. J. Epidemiol.* 144, 421–424. doi: 10.1093/oxfordjournals.aje.a008944

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The psychology of survey response*. London: Cambridge University Press.

Wetzel, E., and Carstensen, C. H. (2014). Reversed thresholds in partial credit models: a reason for collapsing categories. *Assessment* 21, 765–774. doi: 10.1177/1073191114530775

Williams, R., Heinemann, A., Bode, R., Wilson, C., Fann, J., and Tate, D. (2009). Improving measurement properties of the patient health questionnaire-9 with rating scale analysis. *Rehabil. Psychol.* 54, 198–203. doi: 10.1037/a0015529

Appendix

SASS scale on principal's goals for their school.

9. Please indicate how far along you think your school is in -		Mark (X) one box on each line.				
		Just beginning	Long way to go	Almost there	We've reached our goal	Not applicable
a. Implementing educational goals.	0070	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐
b. Implementing organizational/governance goals.	0071	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐
c. Establishing a secure financial base.	0072	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐
d. Attracting and retaining students.	0073	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐
e. Developing a student assessment system.	0074	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐
f. Involving parents in the school.	0075	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐

FORM SASS-2A (7-27-99)

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 06 February 2024
DOI 10.3389/fpsyg.2023.1245863

OPEN ACCESS

EDITED BY
Nikolaos Tsigilis,
Aristotle University of Thessaloniki, Greece

REVIEWED BY
Zhenqiu Lu,
University of Georgia, United States
Michalis Linardakis,
University of Crete, Greece

*CORRESPONDENCE
Fathima Jaffari
✉ f.jaffari@etec.gov.sa

RECEIVED 23 June 2023
ACCEPTED 12 December 2023
PUBLISHED 06 February 2024

CITATION
Jaffari F and Koran J (2024) Model-free measurement of case influence in structural equation modeling.
Front. Psychol. 14:1245863.
doi: 10.3389/fpsyg.2023.1245863

Model-free measurement of case influence in structural equation modeling

Fathima Jaffari[1]*and Jennifer Koran[2]

[1]Department of Tests and Measurement, National Center for Assessment, Education and Training Evaluation Commission (ETEC), Riyadh, Saudi Arabia, [2]Quantitative Methods Program, Southern Illinois University Carbondale, Carbondale, IL, United States

In the field of structural equation modeling (SEM), all commonly used case influence measures are model-based measures whose performance are affected by target-model-misspecification-error. This problem casts light on the need to come up with a model-free measure which avoids the misspecification problem. the main purpose of this study is to introduce a model-free case influence measure, the Deleted- One-Covariance-Residual (*DOCR*), and then evaluating its performance compared to that of Mahalanobis distance (MD) and generalized Cook's distance (*gCD*). The data of this study were simulated under three systematically manipulated conditions: the sample size, the proportion of target cases to non-target cases, and the type of model used to generate the data. The findings suggest that the *DOCR* measure generally performed better than *MD* and *gCD* in identifying the target cases across all simulated conditions. However, the performance of the *DOCR* measure under a small sample size was not satisfactory, and it raised a red flag about the sensitivity of this measure to small sample size. Therefore, researchers and practitioners should only use the *DOCR* measure with a sufficiently large sample size, but not larger than 600.

KEYWORDS

influence analysis, deletion statistics, Mahalanobis distance, generalized Cook's distance, Deleted-One-Covariance-Residual (*DOCR*)

1 Introduction

In structural equation modeling (SEM), normal-distribution-based maximum likelihood (NML) is commonly used as a default estimation method for estimating the parameter values. The NML procedure yields reasonable parameter estimates if the assumption of normality holds in the data distribution. Alternatively, the existence of influential cases in the data might make NML yield biased parameter estimates and affect overall model assessment since these cases could alter the standard error value and the test statistic (Yuan and Bentler, 1998).

One tool that is used for investigating the influence of these cases on the model results is the case influence measures. These measures are built based on the case deletion technique. The case deletion technique is based on the quantification of the impact of the *i*th case by finding the difference between the value of the measure before and after the deletion of the *i*th case to evaluate the impact of this case on the overall model fit. The result obtained from this measure gives information on which case is more influential. In other modeling frameworks, such as OLS regression, there is extensive development and widespread use of case diagnostics for identifying cases, which is not the case with confirmatory factor models, path analysis models, and other models in the SEM framework.

Several regression-based case influence measures have been applied to the SEM field and used with the confirmatory factor models. However, all applied case influence measures are model-based measures that require a theoretical model to be fitted into the data to identify the influential

cases. Because case influence measures are model-based, the accuracy of their performance could be impacted by specification errors (Bollen and Arminger, 1991). Since influence measures rely on the structure of the model, they highlight any case that does not fit the model. The determination of one case fits to the model changes depending on the model that has been fitted to the data. Thus, if the model is misspecified, the case influence measure is expected to yield many cases that cause a poor overall model fit. On the contrary, if the case influence measure reflects a few influential cases, it could be expected that the model was correctly specified, and the actual problem of the influential cases existed among the data (Pek and MacCallum, 2011).

The case influence measures that are commonly used in the field of SEM are all model-based measures. Up to this point, no model-free case influence measure has been proposed in the SEM field. Therefore, the main purpose of this study is to avoid the misspecification problem associated with the performance of model-based measures by developing a model-free case influence measure. The proposed Deleted-One-Covariance-Residual (*DOCR*) measure is based on the covariance matrix of the observed data, which allows the *DOCR* to avoid requiring any specific model to fit the data. The *DOCR* uses the deletion technique by comparing the sample covariance matrix that resulted from deleting the *i*th case from the original sample with the sample covariance matrix that resulted from considering all cases in the original sample $(\mathbf{s_i} - \mathbf{s})$.[2] For standardizing the residuals, the residual difference between the two sample covariance matrices, $(\mathbf{s_i} - \mathbf{s})^2$, is divided by observed variances $(v_m v_j p(p+1))$. After algebraic arranging, the final formula, as seen in Eq. (1), as follows:

$$\boldsymbol{DOCR} = \left[\frac{2}{p(p+1)} \sum_{m=1}^{p} \sum_{j=1}^{m} \frac{(\mathbf{S_i} - \mathbf{S})^2}{v_m v_j} \right] \times 1000 \quad (1)$$

Where $\mathbf{S}$ and $\mathbf{S_i}$ are the sample covariance matrices obtained from original and deleted *i*th case samples, respectively. v_m and v_j are the observed variances of each pair of variables in the covariance matrix, and p is the number of observed variables. Since the *DOCR* measure would otherwise yield small values that range between 10^{-4} and 10^{-5} for the influence of the cases, the formula of this measure includes multiplying by 1000 to make these values more readable. Our goal is to determine whether the purposed model-free measure *DOCR* precisely identifies the influential cases compared to generalized Cook's distance (*gCD*) and Mahalanobis distance (*MD*), which are extensively used in multivariate applications to detect outliers. We present the results of two Monte Carlo simulation studies that compared the performance of the proposed measure to the performance of *MD* and *gCD* in identifying the target cases. We hypothesized that the *DOCR* measure would perform better than *MD* and *gCD* in identifying the target cases across variations in sample size, proportion of target cases, and model specifications.

1.1 Background

In SEM, the case influence measures aim to evaluate the degree of the model fit at the person level; stated differently, they aim to identify unusual cases under the model (Reise and Widaman, 1999). Corresponding to regression, the following factor analysis model as seen in Eq. (2) is considered a latent predictor's multivariate regression model:

$$\mathbf{X_i} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{f_i} + \mathbf{e_i} \quad (2)$$

Where $\boldsymbol{\mu}$ is a population mean vector, $\boldsymbol{\Lambda}$ is a $p \times q$ factor loadings matrix, $\mathbf{f_i}$ is a vector of q-variate latent factors, and $\mathbf{e_i}$ is a vector of measurement errors. Based on this factor model, Yuan and Zhong (2008) stated that the cases with large absolute values of measurement error ($\mathbf{e_i}$) are termed *outliers*, disregarding the values of the factor scores ($\mathbf{f_i}$). The cases with extreme absolute values on the exogenous latent variables' factor scores are termed *leverage cases*. Leverage cases with a small magnitude of measurement errors ($\mathbf{e_i}$) are considered *Good Leverage Cases*, while leverage cases with a large magnitude of measurement errors ($\mathbf{e_i}$) are considered *Bad Leverage Cases*. In SEM, unusual cases with large $\mathbf{e}_i$ are considered influential on both the model fit and the parameters since they cause a large change in the off-diagonal elements of $\mathbf{S}$ (sample covariance matrix). Case influence measures use the deletion technique to quantify the influence of these cases by comparing the value of the statistic before and after the deletion of the *i*th case from the data. Most of these measures have been proposed and developed in the regression field (Belsley et al., 1980; Cook and Weisberg, 1982). However, some of these statistics have been applied to the SEM field to identify the influential cases and quantify their influence on the model findings.

One of the deletion measures that have been applied to SEM is *gCD*. *gCD* is a model-based measure that is used to quantify the influence of the unusual case on the parameter estimates. This measure is a generalized version of Cook's distance (Cook, 1977, 1986). Atkinson (1981) modified Cook's distance for influential case detection by adding the values of the parameter estimates after deleting the *i*th case and controlling for the sample size effect. Then, Lee and Wang (1996) used the generalized least square function to generalize Cook's distance measure to the SEM application.

gCD has been introduced and used in some studies (Zhao and Lee, 1998; Pek and MacCallum, 2011) to examine the case influence on a set of *l* parameters on a set of *l* parameters, as seen in Eq. (3).

$$\boldsymbol{gCD_i} = \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right)' \left[\widehat{\boldsymbol{VAR}}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right)\right]^{-1} \left(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{i}}\right) \quad (3)$$

Where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_i$ are vectors of parameter estimates that are calculated from all cases in the original sample and the sample with the *i*th case deleted, respectively. The $VAR(\hat{\boldsymbol{\theta}}_i)$ is the estimated asymptotic covariance matrix of the parameter estimates calculated from the sample with the *i*th case deleted. Assuming that *k* is the full set of the model parameters and *l* is the number of the desirable subset of the model parameters, one can calculate *gCD* for any subset of parameters *l* instead of the full set of model parameters *k*.

Given the *gCD* quadratic form, the lower bound of *gCD* is equal to zero, which means that this statistic always takes positive values, and that makes *gCD* give us information on the level of change rather than the direction of the change on the model parameters. Thus, a small amount of *gCD* means that a small change in the *l* subset of parameter estimates is associated with the exclusion of the *i*th case from the sample. On the other hand, a large amount of *gCD* means that a large change in the *l* subset of parameter estimates is associated with the exclusion of the *i*th case from the sample.

To obtain information about the direction of change in an individual parameter, the scaled difference $\left(\Delta\hat{\theta}_{ji}\right)$ is used for this specific purpose (Zhao and Lee, 1998; Pek and MacCallum, 2011) as seen in Eq. (4).

$$\Delta\hat{\theta}_{ji} = \frac{\hat{\theta}_j - \hat{\theta}_{j(i)}}{\sqrt{\widehat{\mathbf{VAR}}\left(\hat{\theta}_{j(i)}\right)}} \tag{4}$$

Where $\hat{\theta}_j$ and $\hat{\theta}_{j(i)}$ are the parameter estimates obtained from the original and deleted *i*th samples, respectively. Positive values of difference indicate that small change is associated with the exclusion of the *i*th case and vice versa.

Other case diagnostic measures have been developed for latent variable models (Pek and MacCallum, 2011; Sterba and Pek, 2012). However, these three measures (i.e., *LD*, $\Delta\chi^2$, and *gCD*) are currently the most readily available due to their inclusion in the R package influence. SEM (Pastore and Altoé, 2022).

Due to the slow development of case influence measures in SEM, *MD* is routinely used in multivariate applications to detect unusual cases. *MD* , as seen in Eq. (5), is the distance between the *i*th case and the remaining cases while accounting for the correlation in the data (Mahalanobis, 1936). Some studies used the main and derived versions of this test mainly for detecting the potential multivariate outliers and leveraged cases (Pek and MacCallum, 2011; Yuan and Zhang, 2012).

$$\boldsymbol{MD^2} = \left(\mathbf{Y_i} - \bar{\mathbf{Y}}\right)\mathbf{C^{-1}}\left(\mathbf{Y_i} - \bar{\mathbf{Y}}\right)' \tag{5}$$

$$\boldsymbol{C} = \frac{\mathbf{1}}{\mathbf{n-1}}\left(\boldsymbol{Y_c}\right)'\left(\boldsymbol{Y_c}\right) \tag{6}$$

Where **Y** is an $N \times p$ data matrix containing *N* cases on *p* variables, $\mathbf{Y_i}$ is a $1 \times p$ vector of *p* variables for the *ith* case, $\mathbf{Y_c} = \mathbf{Y} - \bar{\mathbf{Y}}$ is the column-centered data matrix, $\bar{\mathbf{Y}}$ is an $N \times p$ matrix of the column means, $_t$ and and **C**, as seen in Eq. (6), is the variance–covariance matrix (De Maesschalck et al., 2000, p.2). MD^2 distributes as a central chi-square distribution with degrees of freedom (*df*) equal to the number of variables. A significantly low value of p of high MD_i^2 in the corresponding $\chi^2(df)$ means that the *i*th case is a potential outlier (Kline, 2016, p. 73).

However, *MD* is a model-free measure of outlying status rather than case influence, and it is generally used in multivariate applications to detect outliers (Mahalanobis, 1936). In practice, some researchers use *MD* to identify the outliers and delete them prior to fitting the model to the data. The problem with this practice is that influential cases could be outlying cases (i.e., outliers), but not all outlying cases are influential. That is, some outlying cases are not regression outliers because they do not deviate from the linear pattern of the data, so they are considered good cases since their inclusion in the estimation process could lead to a better overall model fit and precise parameter estimates (Rousseeuw and van Zomeren, 1990). Based on this fact, Pek and MacCallum (2011) recommended against using such practice since the removal of good cases, because *MD* identifies them as outlying cases, might lead to worsening the overall model fit. Thus, this practice sheds light on the limitations of using *MD* in the case influence analysis to identify influential cases. On the contrary, model-based measures demand to fit a theoretical model to the data for quantifying the impact that each case exerts on the findings of modeling. The latter measures consider the structure of the model, and their values change as the model structure and set of independent variables change (Belsley et al., 1980).

The purpose of this study is to introduce a model-free case influence measurement that overcomes the problem of specification error and the limitations of using an outlying status measure (i.e., *MD*) in identifying the influential cases. This proposed measure is compared to *MD* and *gCD* to evaluate its ability to identify target cases under a variety of systematically manipulated conditions while accounting for sampling variability using Monte Carlo simulation.

2 Methods

2.1 Data generation

2.1.1 Simulation study 1

The data for this simulation study were generated under a population confirmatory factor analysis (CFA) model with two factors and three indicators per factor. For scaling the factors, the unit variance identification method was used. Target cases were generated from a $N\left(0, 2.25\boldsymbol{I_6}\right)$ distribution (c.f., Lee and Wang, 1996), where $\boldsymbol{I_6}$ is a 6×6 identity matrix. Non-target cases were generated using the common factor model $N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Phi\Lambda}' + \boldsymbol{\Psi}$ is the 6×6 population covariance matrix, $\boldsymbol{\Lambda}$ is the loading matrix with $\boldsymbol{\Lambda}' = \begin{bmatrix} 0.8 & 0.8 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0.5 & 0.5 \end{bmatrix}$, $\boldsymbol{\Phi} = \begin{bmatrix} 1.0 & 0.6 \\ 0.6 & 1.0 \end{bmatrix}$ is the factor correlation matrix, and $\boldsymbol{\Psi} = diag\left\{0.36, 0.36, 0.36, 0.75, 0.75, 0.75\right\}$ is the 6×6 diagonal matrix of unique variances.

2.1.2 Simulation study 2

The data for this simulation study were generated under a population path model with five observed variables. Data sets were simulated with target cases from $a\,N\left(\mathbf{0}, 6.49\boldsymbol{I_5}\right)$ distribution, where $\boldsymbol{I_5}$ is the 5×5 identity matrix, and 6.49 was the result of multiplying the largest variance in the diagonal of the covariance matrix of the data by 4 following the same process of generating the target cases used within the first simulation study (c.f., Lee and Wang, 1996). Non-target cases were generated using the population path model from $a\,N\left(\mathbf{0}, \boldsymbol{\Sigma}\right)$, where

$\mathbf{Y} = \boldsymbol{\Gamma}\mathbf{X} + \mathbf{B}\mathbf{Y} + \boldsymbol{\zeta}$, $\boldsymbol{\Gamma} = \begin{bmatrix} 0.7 & 0.6 \\ 0 & 0.6 \\ 0 & 0 \end{bmatrix}$, $\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.5 & 0.6 & 0 \end{bmatrix}$,

$\boldsymbol{\Sigma}$ is the $v \times v$ population covariance matrix, $\boldsymbol{\Gamma}$ is a parameter matrix of the direct effect of exogenous variables on the endogenous variables, **B** is the parameter matrix of the direct effect of endogenous variables on each other, and $\boldsymbol{\zeta}$ is the matrix of the disturbances.

2.2 Case diagnostics

The *DOCR*, *MD*, and *gCD* were compared. The confirmatory factor analysis models fit in Study 1 are shown in Figures 1, 2. The path analysis models fit in Study 2 are shown in Figures 3, 4. Since model

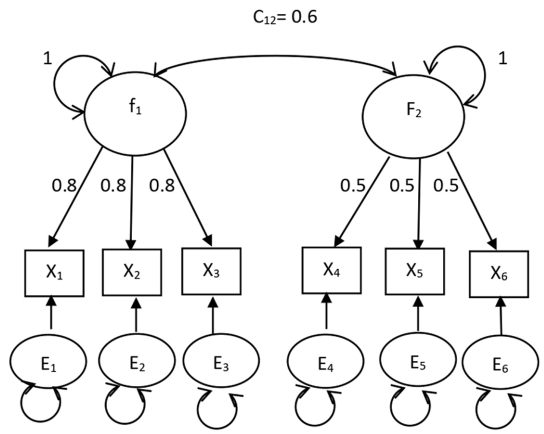


FIGURE 1
The correctly specified common factor model.

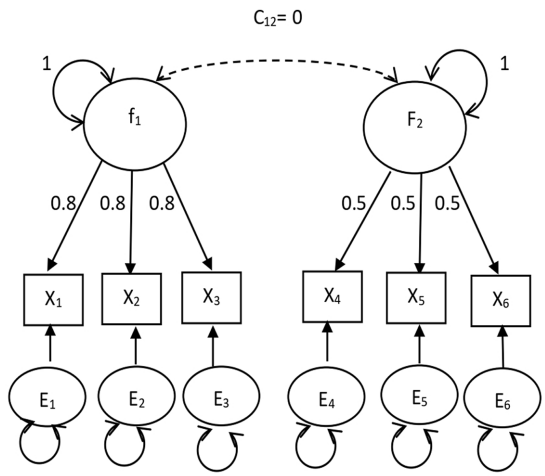


FIGURE 2
The orthogonal common factor model.

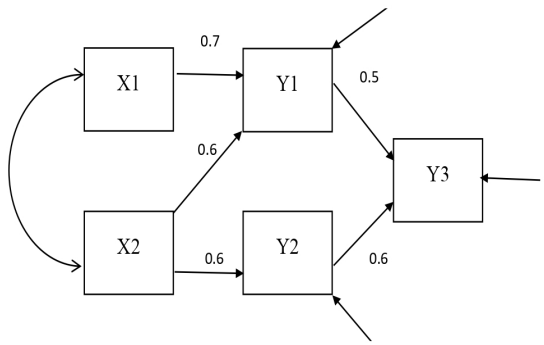


FIGURE 3
The correctly specified path model.

misspecification can affect the identification of target cases, both correctly specified models, shown in Figures 1, 3, and misspecified models, shown in Figures 2, 4, were fit to the simulated data using the

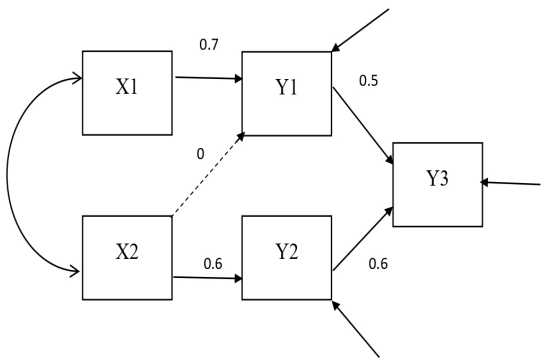


FIGURE 4
The misspecified path model.

R package *lavaan* (Rosseel, 2012). The *DOCR* was calculated using basic matrix functions from the *matlib* package in R (Friendly et al., 2022). The *MD* was calculated using the *mahalanobis* function from the *stats* package that is part of base R. The *gCD* was calculated using the *genCookDist* and *explore.influence* functions from the R package *influence.SEM* (Pastore and Altoé, 2022) for both the correctly specified model and the misspecified models in both studies.

2.3 Implementation

Data were simulated in R v3.4.1 (R Core Team, 2017) with three different sample sizes: 200, 400, and 600. Four proportions of target cases to the number of non-target cases were applied: 0.10, 0.05, 0.02, and 0.01. The sample size and proportion of target cases were fully crossed for a factorial design with 12 conditions. The correctly specified models and misspecified models in both studies were fitted to the data using the R package *lavaan* (Rosseel, 2012). The default boxplot criterion was used to determine cases with high influence (Pastore and Altoé, 2022). The cut-off that determined multivariate outlier cases using *MD* was 12. A preliminary cut-off for *DOCR* was set at 0.01. The miss rate (*MR*) is the ratio of missed target cases to generated target cases, and the false alarm rate (*FAR*) is the ratio of flagged non-target cases to generated non-target cases. Their 95% confidence intervals were computed for each statistic for each replication using R package *psych* (Revelle, 2023). Results were averaged over 100 replications in R with confidence intervals computed using the standard error of the mean and the inverse t distribution. Averages were compared across different statistics and systematic manipulations of the conditions. Example R syntax for computing *DOCR*, as well as *gCD* and *MD*, has been provided in Appendix A. The example in Appendix A has been expanded from the package "influence.SEM" (Pastore and Altoé, 2022).

3 Results

The Study 1 results with the confirmatory factor analysis models are shown in Tables 1, 2, and the Study 2 results with the path analysis models are shown in Tables 3, 4. Tables 1, 3 summarize the miss rates of the three measures, *MD, DOCR*, and *gCD*, by sample size. The *DOCR* measure had the smallest miss rates compared to

TABLE 1 Miss rates for three case detection statistics by proportions of the target to non-target cases for the CFA model.

Prop	0.1		0.05		0.02		0.01	
	$N = 200\ (180 + 20)$		$N = 200\ (190 + 10)$		$N = 200\ (196 + 4)$		$N = 200\ (198 + 2)$	
Statistic	*M*	95%CI	*M*	95%CI	*M*	95%CI	*M*	95%CI
MD	0.429	(0.410, 0.448)	0.340	(0.312, 0.368)	0.353	(0.308, 0.397)	0.280	(0.213, 0.347)
DOCR	0.024	(0.017, 0.029)	0.017	(0.010, 0.024)	0.015	(0.003, 0.027)	0.005	(−0.005, 0.015)
gCD, CS	0.388	(0.368, 0.408)	0.327	(0.298, 0.356)	0.315	(0.270, 0.359)	0.270	(0.208, 0.332)
gCD, MS	0.351	(0.329, 0.371)	0.301	(0.273, 0.329)	0.295	(0.252, 0.338)	0.230	(0.169, 0.291)
	$N = 400\ (360 + 40)$		$N = 400\ (380 + 20)$		$N = 400\ (392 + 8)$		$N = 400\ (396 + 4)$	
MD	0.417	(0.405, 0.429)	0.349	(0.329, 0.367)	0.303	(0.272, 0.333)	0.290	(0.245, 0.335)
DOCR	0.163	(0.151, 0.174)	0.116	(0.103, 0.129)	0.090	(0.071, 0.109)	0.108	(0.077, 0.138)
gCD, CS	0.377	(0.364, 0.391)	0.329	(0.309, 0.348)	0.288	(0.256, 0.322)	0.283	(0.238, 0.327)
gCD, MS	0.348	(0.334, 0.362)	0.297	(0.275, 0.319)	0.275	(0.244, 0.306)	0.275	(0.230, 0.319)
	$N = 600\ (540 + 60)$		$N = 600\ (570 + 30)$		$N = 600\ (588 + 12)$		$N = 600\ (594 + 6)$	
MD	0.422	(0.410, 0.433)	0.342	(0.326, 0.359)	0.331	(0.307, 0.354)	0.288	(0.253, 0.323)
DOCR	0.331	(0.321, 0.340)	0.259	(0.244, 0.274)	0.244	(0.220, 0.268)	0.207	(0.175, 0.238)
gCD, CS	0.385	(0.374, 0.395)	0.322	(0.304, 0.338)	0.318	(0.294, 0.343)	0.293	(0.257, 0.329)
gCD, MS	0.349	(0.339, 0.359)	0.294	(0.277, 0.311)	0.308	(0.285, 0.331)	0.275	(0.237, 0.313)

Prop, proportions of target cases to non-target cases; CI, confidence interval; MD, Malahanobis distance; CS, correctly specified model; gCD, generalized Cook's distance; MS, misspecified model.

TABLE 2 False alarm rates for three case detection statistics by proportions of the target to non-target cases for the CFA model.

Prop	0.1		0.05		0.02		0.01	
	$N = 200\ (180 + 20)$		$N = 200\ (190 + 10)$		$N = 200\ (196 + 4)$		$N = 200\ (198 + 2)$	
Statistic	*M*	95%CI	*M*	95%CI	*M*	95%CI	*M*	95%CI
MD	0.024	(0.022, 0.026)	0.035	(0.033, 0.037)	0.049	(0.046, 0.052)	0.053	(0.051, 0.056)
DOCR	0.701	(0.692, 0.709)	0.759	(0.753, 0.765)	0.804	(0.798, 0.811)	0.820	(0.810, 0.821)
gCD, CS	0.059	(0.056, 0.062)	0.070	(0.066, 0.074)	0.082	(0.079, 0.086)	0.084	(0.079, 0.088)
gCD, MS	0.059	(0.055, 0.062)	0.070	(0.067, 0.074)	0.083	(0.079, 0.087)	0.085	(0.080, 0.089)
	$N = 400\ (360 + 40)$		$N = 400\ (380 + 20)$		$N = 400\ (392 + 8)$		$N = 400\ (396 + 4)$	
MD	0.023	(0.022, 0.025)	0.036	(0.034, 0.037)	0.048	(0.047, 0.049)	0.053	(0.052, 0.055)
DOCR	0.209	(0.205, 0.215)	0.286	(0.282, 0.292)	0.345	(0.341, 0.350)	0.367	(0.362, 0.370)
gCD, CS	0.058	(0.056, 0.061)	0.071	(0.068, 0.073)	0.079	(0.077, 0.082)	0.083	(0.080, 0.085)
gCD, MS	0.057	(0.055, 0.059)	0.072	(0.069, 0.075)	0.081	(0.079, 0.083)	0.084	(0.082, 0.087)
	$N = 600\ (540 + 60)$		$N = 600\ (570 + 30)$		$N = 600\ (588 + 12)$		$N = 600\ (594 + 6)$	
MD	0.023	(0.022, 0.025)	0.036	(0.035, 0.037)	0.050	(0.049, 0.051)	0.053	(0.052, 0.054)
DOCR	0.054	(0.052, 0.056)	0.093	(0.089, 0.095)	0.130	(0.128, 0.133)	0.142	(0.139, 0.144)
gCD, CS	0.057	(0.055, 0.058)	0.071	(0.069, 0.073)	0.079	(0.076, 0.080)	0.082	(0.080, 0.084)
gCD, MS	0.056	(0.054, 0.058)	0.072	(0.070, 0.074)	0.082	(0.079, 0.84)	0.082	(0.080, 0.084)

Prop, proportions of target cases to non-target cases; CI, confidence interval; MD, Malahanobis distance; CS, correctly specified model; gCD, generalized Cook's distance; MS, misspecified model.

MD, *gCD-CS*, and *gCD-MS* for all sample sizes and under the four proportions of the target cases to non-target cases. In addition, the miss rate of the *DOCR* increased significantly as the sample size increased from 200 to 600 under all proportions of target cases to non-target cases. On the other hand, the miss rates of the *MD* and *gCD* remained the same when the sample size increased from 200 to 600 since their miss rates did not differ significantly with the increase in sample size for all proportions of target cases to non-target cases.

Tables 1, 3 show that the miss rate of the *DOCR* decreased as the proportion of target cases to non-target cases decreased. The *DOCR* measure also showed the same pattern of performance under all proportions of target cases to non-target cases through all sample sizes. Similarly, the *MD* and *gCD* measures showed the same pattern of

TABLE 3 Miss rates for three case detection statistics by proportions of target cases to non-target cases for the path model.

Prop	0.1		0.05		0.02		0.01	
	$N = 200$ (180 + 20)		$N = 200$ (190 + 10)		$N = 200$ (196 + 4)		$N = 200$ (198 + 2)	
Statistic	*M*	95%CI	*M*	95%CI	*M*	95%CI	*M*	95%CI
MD	0.225	(0.207, 0.242)	0.135	(0.114, 0.156)	0.100	(0.071, 0.129)	0.100	(0.060, 0.140)
DOCR	0.013	(0.008, 0.017)	0.005	(0.001, 0.009)	0.005	(0.002, 0.012)	0.000	(0.000, 0.000)
gCD, CS	0.103	(0.090, 0.116)	0.075	(0.058, 0.092)	0.075	(0.050, 0.099)	0.075	(0.039, 0.111)
gCD, MS	0.088	(0.076, 0.100)	0.060	(0.044, 0.076)	0.065	(0.041, 0.089)	0.050	(0.020, 0.080)
	$N = 400$ (360 + 40)		$N = 400$ (380 + 20)		$N = 400$ (392 + 8)		$N = 400$ (396 + 4)	
MD	0.229	(0.218, 0.242)	0.136	(0.121, 0.150)	0.095	(0.076, 0.114)	0.090	(0.063, 0.117)
DOCR	0.070	(0.063, 0.077)	0.034	(0.026, 0.041)	0.033	(0.019, 0.046)	0.025	(0.010, 0.040)
gCD, CS	0.101	(0.091, 0.111)	0.069	(0.058, 0.079)	0.066	(0.049, 0.083)	0.073	(0.045, 0.100)
gCD, MS	0.087	(0.078, 0.096)	0.054	(0.044, 0.063)	0.060	(0.044, 0.076)	0.060	(0.035, 0.086)
	$N = 600$ (540 + 60)		$N = 600$ (570 + 30)		$N = 600$ (588 + 12)		$N = 600$ (594 + 6)	
MD	0.232	(0.222, 0.242)	0.143	(0.131, 0.155)	0.098	(0.082, 0.113)	0.087	(0.063, 0.109)
DOCR	0.157	(0.147, 0.166)	0.089	(0.079, 0.100)	0.058	(0.044, 0.070)	0.033	(0.019, 0.047)
gCD, CS	0.105	(0.095, 0.113)	0.078	(0.069, 0.088)	0.057	(0.045, 0.068)	0.055	(0.037, 0.073)
gCD, MS	0.089	(0.082, 0.098)	0.068	(0.059, 0.077)	0.046	(0.035, 0.056)	0.037	(0.023, 0.050)

Prop, proportions of target cases to non-target cases; CI, confidence interval; MD, Malahanobis distance; CS, correctly specified model; gCD, generalized Cook's distance; MS, misspecified model.

TABLE 4 False alarm rates of three case detection statistics by proportions of target cases to non-target cases for the path model.

Prop	0.1		0.05		0.02		0.01	
	$N = 200$ (180 + 20)		$N = 200$ (190 + 10)		$N = 200$ (196 + 4)		$N = 200$ (198 + 2)	
Statistic	*M*	95%CI	*M*	95%CI	*M*	95%CI	*M*	95%CI
MD	0.003	(0.002, 0.003)	0.007	(0.006, 0.008)	0.017	(0.015, 0.018)	0.024	(0.022, 0.026)
DOCR	0.421	(0.408, 0.433)	0.571	(0.560, 0.581)	0.687	(0.678, 0.697)	0.736	(0.727, 0.744)
gCD, CS	0.066	(0.062, 0.069)	0.086	(0.083, 0.089)	0.095	(0.092, 0.099)	0.095	(0.092, 0.99)
gCD, MS	0.092	(0.089, 0.095)	0.122	(0.118, 0.125)	0.137	(0.134, 0.141)	0.140	(0.136, 0.143)
	$N = 400$ (360 + 40)		$N = 400$ (380 + 20)		$N = 400$ (392 + 8)		$N = 400$ (396 + 4)	
MD	0.002	(0.0017, 0.003)	0.0065	(0.006, 0.007)	0.016	(0.014, 0.017)	0.021	(0.019, 0.023)
DOCR	0.064	(0.061, 0.068)	0.139	(0.133, 0.145)	0.237	(0.231, 0.243)	0.292	(0.285, 0.299)
gCD, CS	0.066	(0.064, 0.069)	0.085	(0.082, 0.088)	0.093	(0.090, 0.096)	0.094	(0.091, 0.096)
gCD, MS	0.092	(0.089, 0.095)	0.121	(0.119, 0.124)	0.137	(0.135, 0.139)	0.141	(0.139, 0.143)
	$N = 600$ (540 + 60)		$N = 600$ (570 + 30)		$N = 600$ (588 + 12)		$N = 600$ (594 + 6)	
MD	0.0018	(0.001, 0.002)	0.005	(0.0045, 0.006)	0.016	(0.014, 0.017)	0.022	(0.021, 0.023)
DOCR	0.011	(0.009, 0.011)	0.033	(0.031, 0.036)	0.079	(0.076, 0.082)	0.112	(0.109, 0.115)
gCD, CS	0.068	(0.066, 0.069)	0.083	(0.082, 0.086)	0.089	(0.087, 0.091)	0.094	(0.093, 0.097)
gCD, MS	0.095	(0.093, 0.097)	0.119	(0.117, 0.120)	0.134	(0.133, 0.137)	0.141	(0.139, 0.143)

Prop, proportions of target cases to non-target cases; CI, confidence interval; MD, Malahanobis distance; CS, correctly specified model; gCD, generalized Cook's distance; MS, misspecified model.

performance under all proportions of target cases to non-target cases. However, the pattern of performance for the three measures (*MD*, *DOCR*, and *gCD*) was not always statistically significant, mainly when the sample size was small.

Tables 2, 4 show the false alarm rates of the three measures, *MD*, *DOCR*, and *gCD*, by sample size. As these tables show, the *DOCR* measure had the highest false alarm rates compared to *MD*, *gCD–CS*, and *gCD–MS* for all sample sizes and under the four proportions of the target cases to non-target cases. Unlike the miss rate, the false alarm rate of the *DOCR* decreased as the sample size increased from 200 to 600 under all four proportions of target cases to non-target cases. In addition, the false alarm rate of the *DOCR* measure differed significantly with the increase in sample size. In other words, there was a significant decrease in the false alarm rate of the *DOCR* measure with the increase in sample size. Conversely, the false alarm rates of the *MD* and *gCD* measures did not change significantly with the increase in sample size.

Tables 2, 4 show that the false alarm rate of the *DOCR* increased as the proportion of target cases to non-target cases decreased. The *DOCR* and *MD* measures reflected the same performance pattern under all four proportions of target cases to non-target cases through all sample sizes. That is, within the same sample size, the false alarm rates of the *DOCR* and *MD* increased significantly as the proportion of the target cases to non-target cases decreased. Similarly, the *gCD* measure reflected the same performance pattern under all proportions of target cases to non-target cases. However, this performance pattern was not always statistically significant, mainly when the small sample size was relatively small.

4 Discussion

This study introduced the *DOCR*, a new model-free case influence measure appropriate for SEM analysis. Two simulation studies compared the performance of the *DOCR* with the performance of two other statistics that may be employed to screen cases in this context. The first was *gCD*, which is a model-based measure of case influence. Like other similar model-based case influence measures, such as likelihood distance and chi-square difference, *gCD* is sensitive to model misspecification. The greater the extent of the model misspecification, the less accurately *gCD* will identify influential cases.

The new *DOCR* statistic was also compared with the performance of *MD*. *MD* is a model-free measure. Thus, it is not sensitive to model misspecification. However, *MD* is a measure of outlying status rather than case influence. Thus, this statistic is less appropriate for detecting cases that will ultimately influence the model results.

The *DOCR* overcomes problems with both of these alternative measures employed to screen cases in SEM analysis. The *DOCR* is model-free. Thus, it is not sensitive to model misspecification. The *DOCR* is also a true case influence measure for SEM analysis, in which the model is fit to the sample covariance matrix. By detecting cases that exert a strong influence on the covariance matrix, the *DOCR* detects cases that will impact the results for the model fit to that covariance matrix.

The results of the two simulation studies suggest that more work is needed to find the optimal cut point for the *DOCR*. The *DOCR* performed better than the other measures in flagging target cases because it recorded the lowest miss rate across all conditions. However, the false alarm rate of the *DOCR* was not reasonable since it incorrectly flagged 42–80% of cases as target cases under a sample size of 200 cases. Although this percentage dropped to 10–30% when the sample size increased, it was still not satisfactory compared to other measures.

With all such measures, there is a compromise between the miss rate and the false alarm rate. Thus, the values of the false alarm rate for the *DOCR* can be made more reasonable by adjusting the cut point to yield a better balance between the miss rate and the false alarm rate. Since establishing a criterion cut point for the *DOCR* measure was outside the scope of this study, it is recommended that future studies establish an optimal cut point criterion for this measure.

The results of the two simulation studies also suggest that the *DOCR* is sensitive to sample size. The *DOCR*'s miss rate increased, and the false alarm rate decreased significantly with an increase in sample size, while the miss rate and false alarm rate of *MD* and *gCD* remained the same. This finding was consistent with previous studies. Previous studies have noted how sample size may affect the performance of case influence measures because the influence of the individual case is weighted by the inverse of the sample size (Pek and MacCallum, 2011). Therefore, a large influence is expected from individual cases in small samples. The findings of this study were consistent with studies that showed the performance of some measures, such as chi-square, that were extremely sensitive to sample size (Boomsma, 1982; Fan et al., 1999). Future studies should investigate methods for reducing the sensitivity of the *DOCR* to sample size.

Given these two limitations, practitioners are recommended not to use the *DOCR* measure with overly small sample sizes (i.e., $N \leq 200$) or overly large sample sizes (i.e., $N > 600$). Instead, practitioners should use the range of sample sizes recommended for SEM studies (Kline, 2016) to obtain the best performance of the *DOCR* measure. Care should be exercised in investigating the cases that are flagged, considering that some of the influential cases identified may be due to sampling variability alone. However, used within these guidelines, the *DOCR* shows promise as a model-free case influence measure appropriate for SEM analysis due to its ability to overcome the limitations of existing measures. Example R syntax for computing *DOCR* has been provided in the Appendix.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

FJ and JK designed the study and created the routine. FJ contributed to the write-up of the manuscript, the R code for the DOCR and other indices, the analysis of the data, and summarizing the results. JK contributed to the write-up of the manuscript and to the improvement of all sections of this manuscript. All authors contributed to the article and approved the submitted version.

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.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1245863/full#supplementary-material

References

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika* 68, 13–20. doi: 10.1093/biomet/68.1.13

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Vol. 571*. Hoboken, New Jersey: John Wiley & Sons.

Bollen, K. A., and Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociol. Methodol.* 21, 235–262. doi: 10.2307/270937

Boomsma, A. (1982). "The robustness of LISREL against small sample sizes in factor analysis models" in *Systems under Indirect Observation: Causality, Structure, Prediction*. eds. K. G. Jöreskog and H. Wold (Amsterdam: North-Holland), 149–173.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics* 19, 15–18. doi: 10.1080/00401706.1977.10489493

Cook, R. D. (1986). Assessment of local influence. *J. R. Stat. Soc. Series B* 48, 133–155. doi: 10.1111/j.2517-6161.1986.tb01398.x

Cook, R. Dennis, and Weisberg, Sanford. *Residuals and Influence in Regression*. New York: Chapman and Hall, (1982).

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L. (2000). The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* 50:18. doi: 10.1016/S0169-7439(99)00047-7

Fan, X., Thompson, B., and Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Struct. Equ. Model. Multidiscip. J.* 6, 56–83. doi: 10.1080/10705519909540119

Friendly, M., Fox, J., Chalmers, P., Monette, G., and Sanchez, G. (2022). Matlib: Matrix Functions for Teaching and Learning Linear Algebra and Multivariate Statistics. R package version 0.9.6.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling (4th ed.)*. New York, NY: The Guilford Press.

Lee, S.-Y., and Wang, S.-J. (1996). Sensitivity analysis of structural equation models. *Psychometrika* 61, 93–108. doi: 10.1007/BF02296960

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proc. Natl Inst. Sci. India* 2, 49–55.

Pastore, M., and Altoé, G. (2022). Influence. SEM: Case Influence in Structural Equation Models. R package version 2.3. Available at: https://CRAN.R-project.org/package=influence.SEM

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 (2017). R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/

Reise, S. P., and Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: a comparison of item response theory and covariance structure approaches. *Psychol. Methods* 4, 3–21. doi: 10.1037/1082-989X.4.1.3

Revelle, W. (2023). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.3.12. Available at: https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss. v048

Rousseeuw, P. J., and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* 85, 633–639. doi: 10.1080/01621459.1990.10474920

Sterba, S. K., and Pek, J. (2012). Individual influence on model selection. *Psychol. Methods* 17, 582–599. doi: 10.1037/a0029253

Yuan, K.-H., and Bentler, P. M. (1998). Structural equation modeling with robust covariances. *Sociol. Methodol.* 28, 363–396. doi: 10.1111/0081-1750.00052

Yuan, K.-H., and Zhang, Z. (2012). Structural equation modeling diagnostics using R package semdiag and EQS. *Struct. Equ. Model. Multidiscip. J.* 19, 683–702. doi: 10.1080/10705511.2012.713282

Yuan, K.-H., and Zhong, X. (2008). 8. Outliers, leverage observations, and influential cases in factor analysis: using robust procedures to minimize their effect. *Sociol. Methodol.* 38, 329–368. doi: 10.1111/j.1467-9531.2008.00198.x

Zhao, Y., and Lee, A. H. (1998). Theory and methods: influence diagnostics for simultaneous equations models. *Aust. N. Z. J. Stat.* 40, 345–358. doi: 10.1111/1467-842X.00038

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 26 February 2024
DOI 10.3389/fpsyg.2024.1354072

OPEN ACCESS

EDITED BY
Iasonas Lamprianou,
University of Cyprus, Cyprus

REVIEWED BY
Dimitrios Stamovlasis,
Aristotle University of Thessaloniki, Greece
Aikaterini Vasiou,
University of Crete, Greece

*CORRESPONDENCE
Faye Antoniou
✉ fayeantoniou@eds.uoa.gr

RECEIVED 11 December 2023
ACCEPTED 11 January 2024
PUBLISHED 26 February 2024

CITATION
Antoniou F, Alghamdi MH and Kawai K (2024) The effect of school size and class size on school preparedness.
Front. Psychol. 15:1354072.
doi: 10.3389/fpsyg.2024.1354072

The effect of school size and class size on school preparedness

Faye Antoniou[1]*, Mohammed H. Alghamdi[2] and Kosuke Kawai[3]

[1]Department of Educational Studies, National and Kapodistrian University of Athens, Athens, Greece, [2]Department of Self-Development Skills, King Saud University, Riyadh, Saudi Arabia, [3]David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, United States

The purpose of the present study was to understand students' school readiness as a function of student and teacher behaviors but also school size and class size using both linear and non-linear analytical approaches. Data came from 21,903 schools distributed across 80 countries as per the 2018 cohort of the PISA database. Results pointed to a preference for the Cusp model in that the relationship between school and class sizes with achievement proved to be best described by the non-linearity of the Cusp catastrophe model. The critical benchmarks were a school size of 801 students and a class size of 27 students for which increases beyond those thresholds were linked to nonlinearity and unpredictability in school readiness. For this reason, we suggest using the cusp catastrophe model from Nonlinear Dynamical Systems Theory (NDST) to understand more fully such complex phenomena.

KEYWORDS

cusp catastrophe model, NLDPS, school readiness, PISA 2018, school size, class size

1 The effect of school size and class size on school preparedness

It's important to note that a school's size and course offerings greatly affect academic performance. A smaller class size allows the instructor to deliver a more tailored education, which increases the likelihood of meeting each student's requirements and concerns. This may create a learning environment where students feel recognized and understood, which will boost their engagement and drive (Ajami and Akinyele, 2014). Teachers in smaller courses are also better at using a variety of instructional strategies, accommodating different learning modalities, and creating dynamic and interesting learning environments (Bradley and Taylor, 1998). Since there are fewer students to watch in a classroom, keeping discipline is easier, which may reduce disturbances and improve learning. This is shown by greater results on standardized examinations as well as more positive long-term educational outcomes (Mosteller, 1995; Tseng, 2010; Krassel and Heinesen, 2014; Gereshenson and Langbein, 2015; Lowenthal et al., 2019). According to Blatchford et al. (2011), Nandrup (2016), and Yamamori et al. (2021), the number of pupils in a class has a considerable impact on both the educational experience students have and the academic results they attain.

Class size and the ratio of staff to students are often used as measures to evaluate the quality of higher education (Thom, 1975; Molenaar and Oppenheimer, 1985; Brown, 1995; Bandiera et al., 2010; Martin 2015; Konstantopoulos and Shen, 2023). These studies were conducted by Stewart and Peregoy, 1983; Guastello, 1984, 1987,1992; Bandiera et al. (2010); Konstantopoulos and Shen (2023), and Martin (2015). Some studies suggest that larger class sizes hurt student learning; however, a significant number of studies present findings that are inconclusive or demonstrate a combination of positive and negative effects (Bellante, 1972;

Edgell, 1981; Hancock, 1996; Kennedy and Siegfried, 1997; Hill, 1998; Jarvis, 2007; Gleason, 2012; De Paola et al., 2013; Matta et al., 2015; Olson et al., 2011). However, previous research has shown that students tend to have a more positive perception of their learning experience when the number of classes they are required to attend is decreased (Kwan, 1999; Van der Maas and Molennar, 1992; Bedard and Kuhn, 2008; Westerlund, 2008; Mandel and Sussmuth, 2011; Monks and Schmidt, 2011; Benton and Cashin, 2012; Sapelli and Illanes, 2016; Hufford et al., 2003; Stamovlasis, 2006). The use of active learning strategies by teachers in smaller class sizes, in addition to the provision of more individualized attention to students (Lammers and Murphy, 2002; Arias and Walker, 2004; Kokkelenberg et al., 2006; Monks and Schmidt, 2011; Van der Maas et al., 2003), could be the reason for this phenomenon (Lammers and Murphy, 2002; Arias and Walker, 2004; Kokkelenberg et al., 2006; Monks and Schmidt, 2011). However, the amount of material that is currently available about the challenges that are related to the application of active learning approaches in smaller class sizes is quite limited (Wright et al., 2017).

A large, influential study namely, the Student Teacher Achievement Ratio (STAR) study reported significant benefits from class size reductions on students' achievement, if these reductions take place early with the effects being more pronounced for students from disadvantaged family backgrounds (Word et al., 1990; Mosteller, 1995; Finn and Achilles, 1999; Krueger, 1999; Nye et al., 2000). These findings are backed up by several academic publications, such as those written by Guastello (2001), Word et al. (1990), Mosteller (1995), Finn and Achilles (1999), Krueger (1999), and, Nye et al. (2000), amongst others. Throughout the early years of their schooling, policymakers need to reflect on the most effective way to direct Corporate Social Responsibility (CSR) programs toward children who are starting in life with socioeconomic disadvantages. Policymakers also have the option of choosing to offer funding for CSR programs while at the same time providing local school leaders the liberty to decide how such programs will be implemented. It is essential to take into consideration a cost–benefit analysis of educational policy whenever one is charged with making judgments about the maximum number of students allowed in a given classroom.

1.1 Evaluating the type of relationship between school size, class size, and school outcomes

Past studies have primarily engaged linear modeling to evaluate the role of class size on school outcomes. The idea that the relationship between class size and student, teacher, and school outcomes is linear falls short for the following reasons. Linear models operate under the assumption that the relationship between variables is best depicted by a mathematical straight line. However, human behavior and outcomes, such as learning and teaching, are inherently complex and multi-faceted. A simple linear relationship might not capture all the nuances and intricacies involved.

Empirical evidence has reported both linear and non-linear effects using, for example, quadratic models. The problem with those findings is that they reported both positive and negative nonlinear trajectories that contradict each other (e.g., Foreman-Peck and Foreman-Peck, 2006; Crispin, 2016). It is possible, however, that the impact of class size on outcomes changes when a certain threshold is crossed. For example, reducing a class from 40 to 30 students might have a more significant impact than reducing it from 30 to 20. Such an effect was reported in the Lee and Smith (1997) study as they reported that school sizes ranging in number of students from 600 and 900 were optimal in facilitating reading and math outcomes. Linear models would most likely be ineffective in capturing such non-linear effects. For the above, more elaborate, and complex analytical models that take into account non-linear changes in behavior and operate using multiple predictors who may exert linear and nonlinear effects are needed. Empirical evidence of this effect has been provided by Cobb et al., 1983; Cobb and Zacks, 1985; Bowne et al. (2017) who reported that "both class size and child-teacher ratio showed nonlinear relationships with cognitive and achievement effect sizes" (p. 407) with implications that this relationship may be further complicated by SES associations. In a large-scale study by Cobb, 1981; Lee and Loeb (2000), concerns were raised about the hypothesized relationship between school size and student learning. In the present study, we propose that the relationship between class size and school outcomes becomes non-linear following a crucial threshold in class size, beyond which, student and school outcomes become unpredictable and chaotic (Oliva et al., 1987). This is why the cusp model may provide the most appropriate means to evaluate the proposed relationships.

1.2 Nonlinear dynamical systems theory and the cusp catastrophe

The cusp catastrophe model is essential in nonlinear dynamic systems theory for explaining sudden and discontinuous system state changes generated by continuous variables. The model works across fields with more recent applications in psychology, education, medicine, and public health. It efficiently handles complex linear and nonlinear interactions between independent variables (Chen and Chen, 2017) toward the understanding of behavioral phenomena. As control variables change, the cusp model can decipher abrupt and sudden behavioral changes offering insights into the stability and transitioning of human behavior (Chen et al., 2014). Applications in the field of education include the investigation of cognitive functioning (Stamovlasis, 2011; Tsitsipis et al., 2012), the teaching of physics (Papageorgiou et al., 2010), and chemistry (Stamovlasis et al., 2005). In the examination of stress and trauma the cusp model has contributed significantly to our understanding of changes in human emotions as individuals transition from one psychological state to another (Kira et al., 2019, 2020). Thus, the cusp catastrophe model has recently been popularized in the social sciences to examine scenarios where shuttle changes in a parameter are associated with drastic and dramatic changes in outcome variables The model employs a potential function, $f(y; a, b)$ for a single dependent variable y given linear and nonlinear parameters a and b:

$$f(y; a, b) = ay + 1/2by^2 - 1/4y^4 \text{ (Eq. 1).}$$

The phenomenon under study is termed the "catastrophe set" which evaluates outcomes inside the parameter space of coordinates (a, b). As shown in Figure 1, when the "b" terms (bifurcation variables), in our instance school size and class size are at low levels changes in school readiness are expected to be linear and smooth, likely fitting the premises of the linear model (see expected Pattern A to linearity). When, however, increases in school size and/or class size exceed a worrisome, critical level, school readiness oscillates between two behavioral modes, low and high

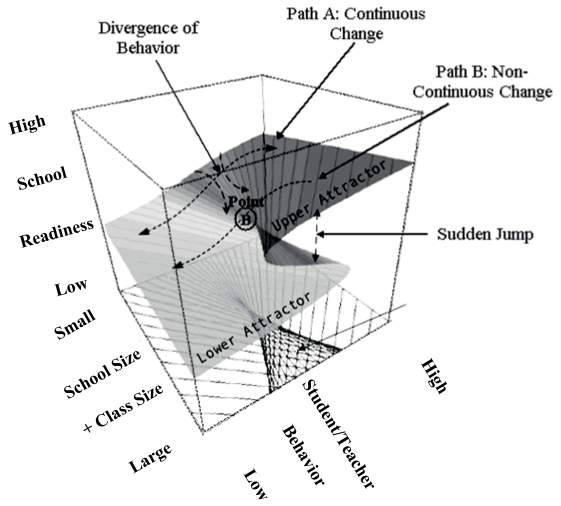


FIGURE 1
Description of the cusp model with the outcome variable school readiness; it was predicted linearly using the asymmetry terms of student and teacher behaviors and non-linearly using the bifurcation terms of school size and class size (Tasker, 2003).

readiness, reaching a state of unpredictability (see Pattern B to non-linearity). Point B in the figure is termed "the cusp point" and reflects unpredictable changes in the outcome variable following midpoint levels in the bifurcation variable. In other words when schools and class sizes are small and move toward medium levels, school readiness is expected to covary in a linear manner; this linear prediction is disturbed following some critical levels of both school size and class size suggesting that increases in these variables are no longer adaptive.

The purpose of the present study was to understand students' school readiness as a function of student and teacher behaviors but also school size and class size using both linear and non-linear data analysis procedures to understand more fully such complex relations.

2 Method

2.1 Participants and procedures

Data came from 21,903 schools distributed across 80 countries as per the 2018 cohort of the PISA database (OECD, 2019). The unit of analysis was the school. Thus, student and teacher estimates were aggregated per school. Student participants in PISA 2018 had to be in the range of 15 years 3 months and 16 years and 2 months and had to be in grade 7 or above. Procedures regarding ethics and sample selection are described here (https://www.oecd.org/pisa/publications/pisa-2018-results.htm). Data may be accessed at https://www.oecd.org/pisa/data/.

2.2 Measures

All measures were derived from the PISA 2018 most recent cohort.

2.2.1 School readiness

This scale is comprised of 8 items completed by school principals on factors that hinder a school's capacity to provide instruction. The items use the stem: "Is your school's capacity to provide instruction hindered by any of the following issues" with the content relating to the shortage of teaching staff, insufficient instructional materials, inadequate physical infrastructure, poor lab equipment, and shortage of ICT resources (Werblow and Duesbery, 2009; Wobmann and West, 2006). Scaling ranged between "not at all" and "a lot" (see Supplementary Appendix A1). Given the high internal consistency reliability of the items with omega being at 0.839, factor scores were estimated using maximum likelihood.

2.2.2 Student behavior hindering learning

It was assessed using 3 student-reported items that evaluate how often students disrupt lessons (a) with noise and talking, (b) with misbehavior, and (c) by being late or absent (Alexander et al., 1992). The items are scored using a 4-point scaling system anchored between "never" and "almost every day." Scores were estimated using Weighted Likelihood Estimation (WLE).

2.2.3 Teacher behavior hindering learning

A 4-item scale was created using student responses to assess how often the teacher (a) explains things in a way that is difficult to understand, (b) does not give enough help when you need it, (c) does not seem to care about whether or not you learn, and (d) does not keep order in the classroom. Items engage a 4-point scaling system from "never" to "almost every day." Estimated factor scores utilized the WLE estimator.

2.2.4 Class size and school size

Class size reflected the number of students in the class using a categorically ordered variable with intervals of 5 students. The categories ranged between "fewer than 15 students" to "more than 50 students." School size was measured as a summative variable expressing the number of students in the school. It is a measure of total enrollment rather than expressing estimates for particular grades, cohorts, genders, or else.

2.3 Statistical data analyses

2.3.1 Cusp catastrophe

At present, there exist multiple analytical models that can be utilized for the detection of a cusp catastrophe. Among these models, Cobb's (1998) methodology and its implementation in R using Grasman's cusp package (Grasman et al., 2009) are widely recognized and accepted. Additionally, the polynomial regression model proposed by Guastello (2002) and the modifications made by Chen et al. (2020) to Cobb's method are also noteworthy alternatives although limited by the unavailability of routines in statistical packages. While both Cobb's and Guastello's models have been widely used, we consider Cobb's (1981) method as more closely associated with catastrophe theory whereas Guastello's methodology is more general and includes various forms of non-linear regression models, such as the quadratic. Thus, we choose the methodology proposed by Cobb (1998), which gained popularity through its implementation in R. To achieve optimality, several conditions must be satisfied by the model. Firstly, the asymmetry and bifurcation parameters should exhibit a significant effect. Secondly, the cusp model should demonstrate superiority over linear and non-linear competing models such as the logistic, the

quadratic, and the cusp. Third, a relatively small proportion of observations, approximately 10%, should fall within the bifurcation area. Lastly, there should be evidence indicating the presence of bimodality within the bifurcation area and multimodality elsewhere for the outcome variable. We deferred using the pseudo-R-squared statistic provided by the package given that it can take negative values. Of note here is the sign of the bifurcation term(s) which requires additional elaboration. Assuming that the bifurcation term is scaled so that higher scores are indicative of unpredictability, then a positive coefficient should be observed. A positive coefficient indicates that the relationship between bifurcation and the outcome variable is linear at low levels of the bifurcation term and becomes non-linear later on. Based on that, in the present study, a positive slope in the bifurcation term was desirable. On the contrary, a negative bifurcation term suggests that at low levels of the splitting factor/bifurcation term, the system is chaotic, and as the scores on the splitting factor increase, linearity, and equilibrium are gradually present. This is not the case in the present study for which unpredictability is expected when increases in school and class size move beyond some adaptive level.

3 Results

3.1 Prerequisite assumptions of cusp catastrophe

One of the important assumptions of the cusp catastrophe is that the dependent variable must have more than one mode. In the present study, the latent school readiness variable presented itself with multimodality as shown in Figure 2 satisfying the prerequisites of the cusp model. Further evidence of the multimodality of school readiness is shown in Figure 3 using the mode tree (Minnotte and Scott, 1993). The figure displays the mode locations of the readiness variable for each bandwidth. The figure displays 25 modes suggesting multimodality as does Figure 2.

3.2 Predicting school readiness from student and teacher behaviors linearly and from school and class size nonlinearly

Intercept and slope terms of the cusp model are shown in Table 1 with all terms being significant. School readiness was positively predicted by the linear contribution of student and teacher behaviors ($b_{\text{Student}} = 0.196$, $p < 0.001$; $b_{\text{Teacher}} = 0.264$, $p < 0.001$). Interestingly, both bifurcation terms were also significant signaling non-linearity. Specifically, as school size and class size increase beyond a specific critical threshold, their relationship to school readiness becomes chaotic and unpredictable ($b_{\text{School size}} = 0.001$, $p < 0.001$; $b_{\text{Class size}} = 0.004$, $p = 0.027$). The critical benchmarks were a school size of 801 students and a class size of 27 students for which increases beyond those thresholds were linked to non-linearity and unpredictability in school readiness. When testing optimal model fit between linear and non-linear models (see Table 2) results pointed to the superiority of the cusp model using the information criteria values of AIC, AICc, and BIC over the competing models (a) linear, (b) quadratic, and (c) logistic, which were consistently lower in the cusp model compared to all other models. Furthermore, a chi-square test contrasted linear and cusp models pointed to a significant misfit of the linear model [$\chi^2(2) = 1732$, $p < 0.001$]. Further evidence for the cusp model's preference is shown in Figure 4, with multimodal distributions at various areas across the response surface and bimodality within the bifurcation area (bottom right figure). Visually speaking, Figure 5 displays the observations as they oscillate between upper and lower surfaces and within the bifurcation area again fitting the expectations of the cusp catastrophe model. An ancillary to Figure 5, is Figure 6, which displays the relative position of the observations to the upper and lower surfaces using a control plane scatterplot. Observations with darker colors (e.g., purple) are positioned closer to the upper surface and those with lighter colors (e.g., light green) are closer to the lower surface (see Grasman et al., 2009).

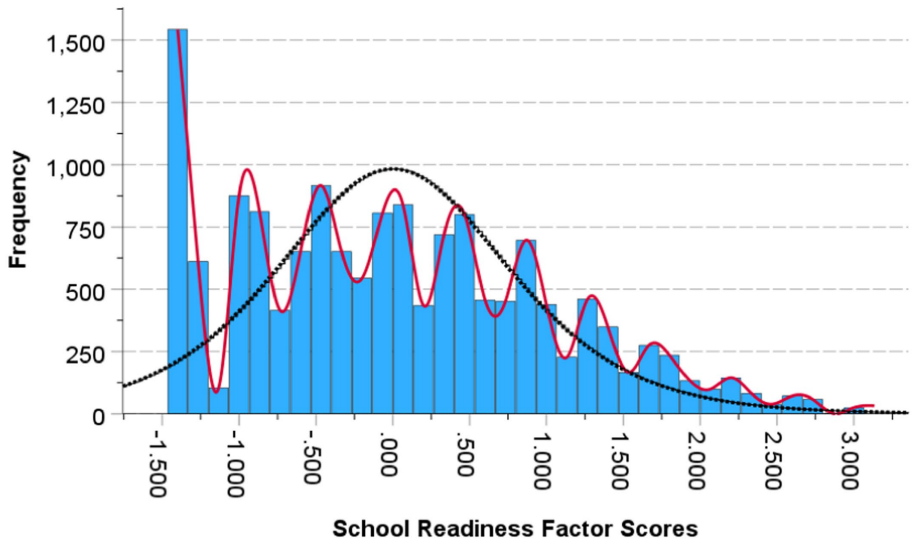


FIGURE 2
Plotting the multimodal distribution of the dependent variable school readiness.

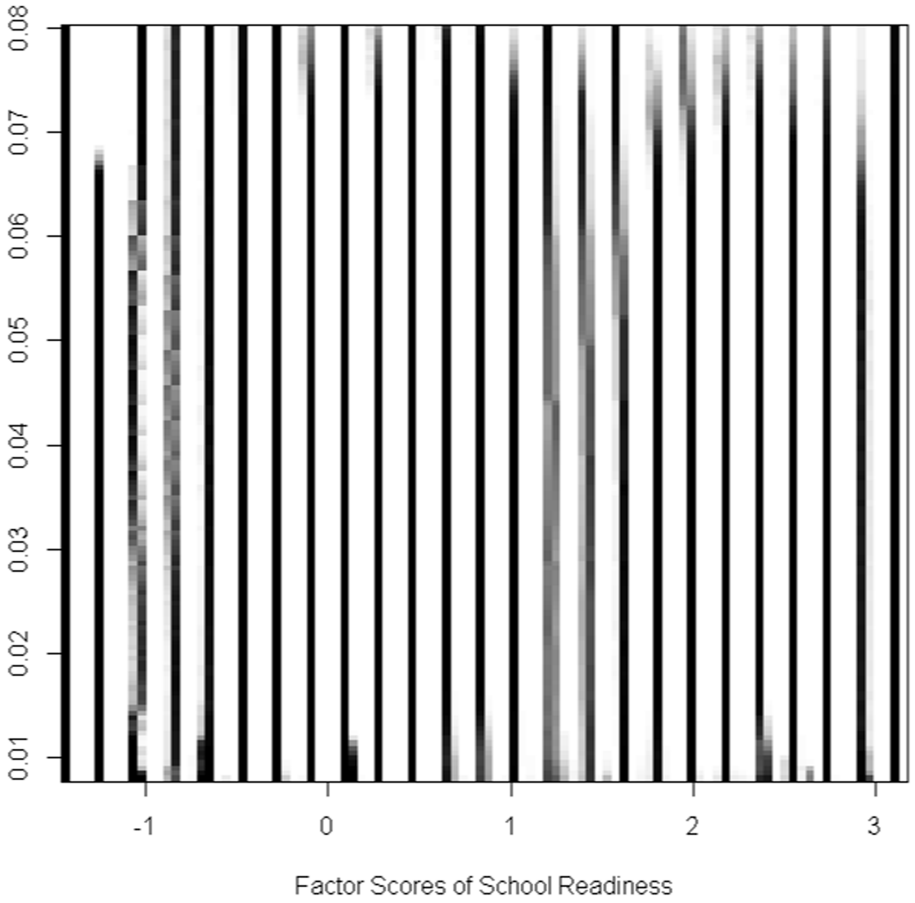


FIGURE 3
Mode locations for each bandwidth of the school readiness factor scores construct. The number of modes reflects those shown in Figure 2.

TABLE 1 Parameter estimates of the cusp model for the prediction of school readiness using a combination of asymmetry (student and teacher behaviors) and bifurcation (school and class size) predictors.

Cusp model intercept and slope terms	Unstandardized B	S.E.	*Z*-test	Value of *p*
a (Intercept)	−0.870	0.027	−31.986	<0.001***
a_1 (Student behavior)	0.196	0.012	15.911	<0.001***
a_2 (Teacher behavior)	0.264	0.019	13.803	<0.001***
b (Intercept)	−0.421	0.077	−5.468	<0.001***
b_1 (School size)	0.001	0.001	18.933	<0.001***
b_2 (Class size)	0.004	0.001	2.212	0.027*
w (Intercept)	−0.496	0.009	−51.496	<0.001***
w (School readiness)	0.770	0.004	173.993	<0.001***

Note: The terms a, b, and w refer to asymmetry, bifurcation, and outcome variables' intercept terms with those including a subscript the specific slopes for the asymmetry and bifurcation variables. ***p <. 001; **p < 0.01; *p < 0.05.

4 Discussion

The purpose of the present study was to understand students' school readiness as a function of student and teacher behaviors but also school size and class size using both linear and non-linear analytical approaches. Results pointed to a preference for the cusp catastrophe model in that the relationship between school and class size with achievement is determined by specific thresholds of these variables.

Past research indicates that the size of a class has a significant influence on the academic achievement of students. The work by Kenayathulla et al. (2019) favored the role and functioning of smaller classrooms and their positive impact on academic achievement. Studies on the Portugal Programme Mais Sucesso Escolar (PMSE) indicate that factors such as class size, composition, and tailored instruction might lead to a decrease in grade repetition and an enhancement in academic achievement (Barata et al., 2015).

Nevertheless, these findings have contentious ramifications for educational policy. Other studies on the other hand (e.g., Filges et al., 2018), pointed out that decreasing class sizes has minimal impact and that there are more cost-effective methods for enhancing student achievement. These unequivocal findings demonstrate that class size and outcomes vary based on the circumstances and composition of the student population (Milesi and Gamoran, 2006).

The most important finding of the present study was that preference for the cusp model allowed us to identify important thresholds for which student readiness is no longer predictable. These thresholds were 801 students for school size and 27 students for class size. Interestingly, the estimates for school size agree with earlier suggestions using quadratic models suggesting that between 600 and 900 students is the optimal school size (Lee and Smith, 1997) and also the work of Andrews et al. (2002) who reported dysfunctional schools when exceeding 1,000 students. For class size, earlier work suggested diminishing returns in that reducing class size from 30 to 25 students is more beneficial compared to reducing it from 20 to 15 (Mosteller, 1995; Krueger, 1999). Thus, the currently identified threshold falls within earlier predictions (Word et al., 1990; Finn and Achilles, 1999; Nye et al., 2000).

TABLE 2 Comparing linear and cusp models using information criteria.

Models tested	Loglikelihood	Par	AIC	AICc	BIC
1. Linear	−20651.560	6	41315.11	41315.12	41360.87
2. Logistic	−20579.590	7	41173.18	41173.18	41226.56
3. Quadratic	−20629.389	8	41274.78	41310.37	41335.79
4. Cusp	−19784.772	9	39587.55	39587.56	39656.19

Par, number of freely estimated parameters; AIC, akaike criterion; AICc, corrected AIC; BIC, Bayesian information criterion. Preference is given to the BIC. Its sample-adjusted variant is not included because of the large sample size of the PISA 2018 cohort.

4.1 Study implications for educational policy

The growing body of research suggesting that larger school and class sizes harm student achievement requires a thorough reassessment of educational systems. When class and school sizes become too large, the amount of attention given to each student decreases, which can hinder customized instruction and result in a decrease in academic performance (Blatchford, 2003; Hattie, 2006). This issue highlights the necessity for policymakers to adopt initiatives focused on maintaining or decreasing class and school sizes to cultivate more efficient learning environments. Possible approaches could involve implementing strict class size restrictions, especially in early schooling where personalized attention is vital, and reorganizing bigger educational institutions into smaller learning communities to improve individualization and assistance (Lee and Smith, 1997). Furthermore, it is important to implement laws that provide fair and equal access to small-sized classrooms and schools among all socioeconomic and demographic groups. This will help to resolve any potential inequalities in educational achievements. Furthermore, a transition to smaller educational environments requires corresponding improvements in teacher recruiting, training, and retention methods, guaranteeing that

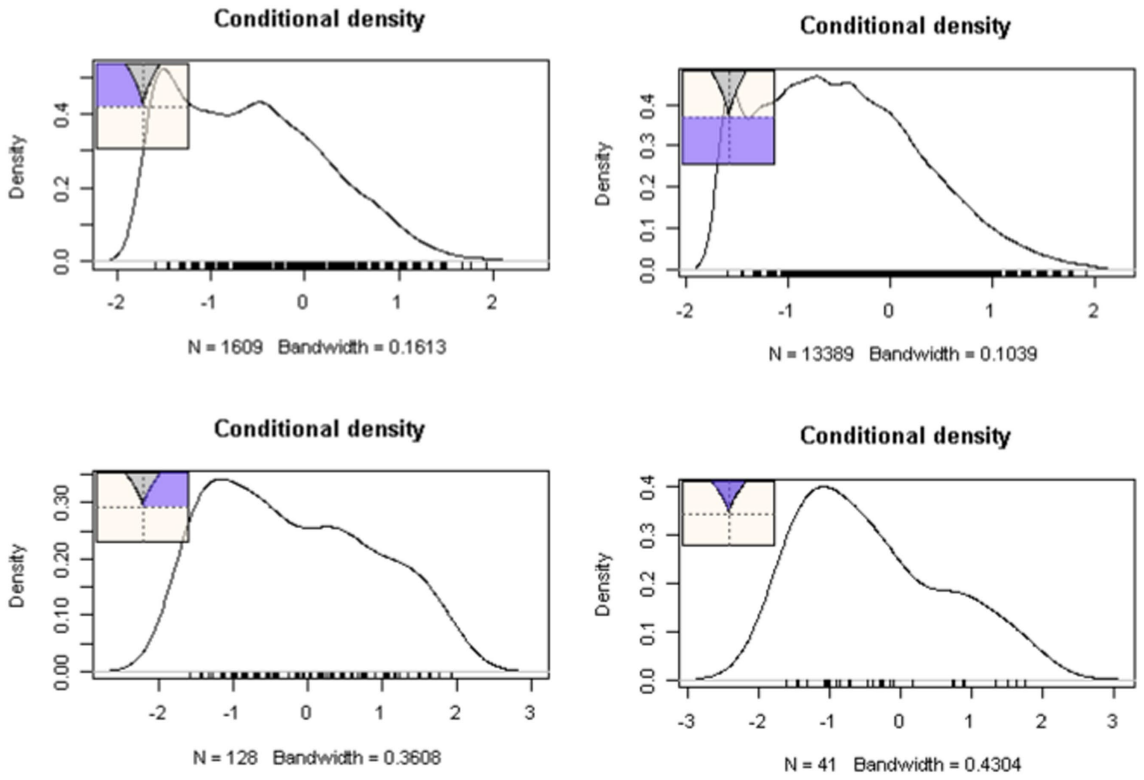


FIGURE 4
Frequency of observations at various places on the response surface as indicated by the shaded areas. The presence of skew and multimodality are apparent in several areas of the lower response surface.

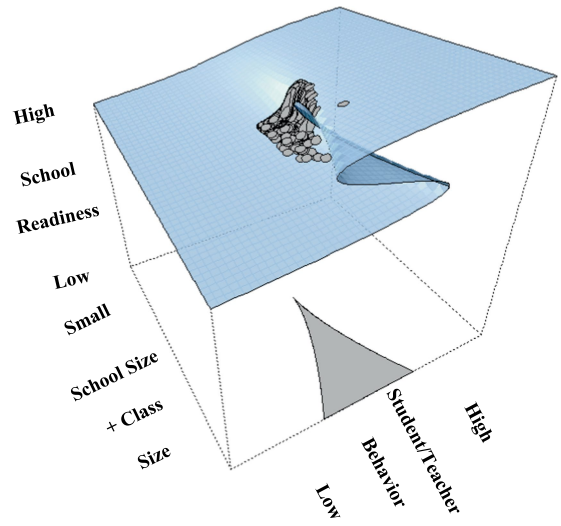


FIGURE 5
Cusp catastrophe upper and lower surfaces with observations transitioning between the two and within the bifurcation area.

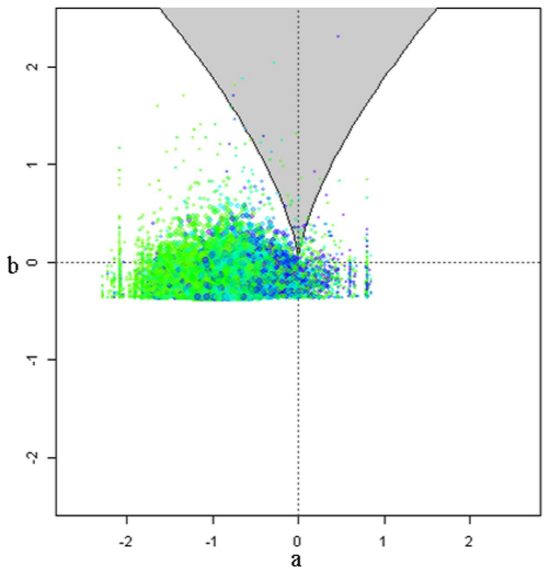


FIGURE 6
Control plane scatterplot with darker dots showing placement of observations close to the upper surface and light color dots towards the lower surface.

the standard of education remains uncompromised. In conclusion, although there are difficulties in managing and funding efforts to optimize class and school sizes, the possibility of achieving substantial enhancements in student performance makes it an essential area of concentration for educational reform and policy formulation.

4.2 Study limitations and future directions

The variables "school size," and "classroom size" have long been recognized as significant in research although past studies have presented several methodological and design deficiencies. The existing research on the functioning of school and classroom size has several limitations. Firstly, it lacks generalizability due to the impracticality of randomly assigning students to schools and classes advising caution before generalizing the present findings beyond the specific sample. Secondly, there is a lack of consistent measures, particularly in quantifying the distinction between "large" and "small" schools with proxy measurements including the number of teachers, the number of students, or other ratio variables, which are also challenged by the violation of distributional assumptions. There is also a need to engage multivariate analyses that can consider the variations in sampling across students and schools using both linear and non-linear means. In the present study, the use of international population data using rigorous procedures for sampling and representativeness in each country overcomes one of the major limitations of past studies. Furthermore, the analytical framework utilized here is not without limitations. The cusp catastrophe model has been criticized as overfitting the data, thus, limiting model generality (Poston and Stewart, 1978). Concerns about the reliability of the findings from small samples have also been raised (Zhang, 2016) as well as the model's generality to real-world phenomena (Schelling, 1973). Others were concerned that the complexity and uncertainty of real-world phenomena cannot be captured by a set of mathematical equations (Borsboom and Cramer, 2013) and specifically one type of asymmetry measured within the cusp catastrophe model (Cramer, 2008).

In the future, we advise the use of the present analytical framework using a per-country analysis as well as the invariance of the findings across important moderating variables such as gender, SES, urbanicity, private or public schooling, and other variables which were found to be important predictors of a school's climate.

5 Conclusion

In this comprehensive examination of the factors influencing students' school readiness, a significant discovery emerged: the relationship between school and class sizes with achievement is best described by the non-linear complexities of the Cusp catastrophe model. This study, utilizing data from over 21,000 schools across 80 countries, revealed critical thresholds at a school size of 801 students and a class size of 27 students. Beyond these points, size increases are associated with unpredictability and decreased school readiness. This suggests a pronounced shift in the traditional understanding of educational environments, emphasizing the importance of maintaining optimal class and school sizes to ensure effective learning and teaching. The findings underscore the need for policymakers to reconsider current educational structures, advocating for more personalized and manageable learning environments to enhance student achievement. While this study provides a groundbreaking insight into the dynamics of educational settings, its reliance on the cusp catastrophe model and the specific thresholds identified necessitates further investigation and validation to ensure widespread applicability and understanding of its implications in the complex landscape of educational policy and practice.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.oecd.org/pisa/data/2018database/.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the patients/ participants or patients/participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

FA: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing. MA: Data curation, Funding acquisition, Investigation, Writing – review & editing. KK: Formal analysis, Methodology, Software, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors extend their appreciation to the Deputyship for Research and Innovation, "Ministry of Education" in Saudi Arabia for funding this research (IFKSUOR3-420-4).

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1354072/full#supplementary-material

References

Ajami, I. R., and Akinyele, O. B. (2014). Effect of student-teacher ratio on academic achievement of selected secondary school students in Port Harcourt Metropolis, Nigeria. *J. Educ. Pract.* 5, 100–106.

Alexander, R. A., Herbert, G. R., DeShon, R. P., and Hanges, P. J. (1992). An examination of least-squares regression modeling of catastrophe theory. *Psychol. Bull.* 111, 366–374. doi: 10.1037/0033-2909.111.2.366

Andrews, M., Duncombe, W., and Yinger, J. (2002). Revisiting economies of size in American education: are we any closer to a consensus? *Econ. Educ. Rev.* 21, 245–262. doi: 10.1016/S0272-7757(01)00006-1

Angrist, J. D., and Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *Q. J. Econ.* 114, 533–575. doi: 10.1162/003355399556061

Arias, J., and Walker, D. (2004). Additional evidence on the relationship between class size and student performance. *J. Eco. Edu.* 35, 311–329.

Bandiera, O., Larcinese, V., and Rasul, I. (2010). Heterogeneous classsize effects: New evidence from a panel of university students. *Econ. J.* 120, 1365–1398.

Bedard, K., and Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Eco. Edu. Rev.* 27, 253–265.

Bellante, D. M. (1972). A summary report on student performance in mass lecture classes of economics. *J. Eco. Edu.* 4, 53–54.

Benton, SL, and Cashin, WE (2012) Student ratings of teaching: A summary of research and literature. IDEApaper #50, Manhattan, KS: The Idea Center.

Blatchford, P. (2003). *The class size debate: Is small better?* Milton Keynes, England: Open University Press.

Blatchford, P., Bassett, P., and Brown, P. A. (2011). Examining the effect of class size on classroom engagement and teacher–pupil interaction: differences in relation to pupil prior attainment and primary vs. secondary schools. *Learn. Instr.* 21, 715–730. doi: 10.1016/j.learninstruc.2011.04.001

Blatchford, P., and Lai, K. C. (2012). "Class size: arguments and evidence" in *International encyclopedia of education*. eds. B. McGraw, E. Baker and P. P. Peterson. *3rd* ed (Oxford, UK: Elsevier)

Borsboom, D., and Cramer, A. O. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* 9, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Bowne, J. B., Magnuson, K. A., Schindler, H. S., Duncan, G. J., and Yoshikawa, H. (2017). A Meta-analysis of class sizes and ratios in early childhood education programs: are thresholds of quality associated with greater impacts on cognitive, achievement, and socioemotional outcomes? *Educ. Eval. Policy Anal.* 39, 407–428. doi: 10.3102/0162373716689489

Bradley, S., and Taylor, J. (1998). The effect of school size on exam performance in secondary schools. *Oxf. Bull. Econ. Stat.* 60, 291–324. doi: 10.1111/1468-0084.00102

Brown, C. (1995). *Chaos and catastrophe theories*. London: Sage.

Chen, D., and Chen, X. (2017). Cusp catastrophe regression and its application in public health and behavioral research. *Int. J. Environ. Res. Public Health* 14:1220. doi: 10.3390/ijerph14101220

Chen, D., Chen, X., Lin, F., Tang, W., Lio, Y., and Guo, Y. (2014). Cusp catastrophe polynomial model: power and sample size estimation. *Open J. Stat.* 4, 803–813. doi: 10.4236/ojs.2014.410076

Chen, X., Wang, K., and Chen, D. G. (2020). "Cusp catastrophe regression analysis of testosterone in bifurcating the age-related changes in PSA, a biomarker for prostate cancer" in *Statistical methods for global health and epidemiology: Principles, methods and applications*. eds. X. Chen and D. G. Chen (Cham, Switzerland: Springer), 353–372.

Cobb, L. (1981). Parameter estimation for the cusp catastrophe model. *Behav. Sci.* 26, 75–78. doi: 10.1002/bs.3830260107

Cobb, L (1998). *An introduction to cusp surface analysis Technical report*, Aetheling Consultants, Louisville, CO, USA.

Cobb, L., Koppstein, P., and Chen, N. H. (1983). Estimation and moment recursion relations for multimodal exponential distributions. *J. Am. Stat. Assoc.* 78, 124–130. doi: 10.1080/01621459.1983.10477940

Cobb, L., and Zacks, S. (1985). Applications of catastrophe theory for statistical modeling in the biosciences. *J. Am. Stat. Assoc.* 80, 793–802. doi: 10.1080/01621459.1985.10478184

Cramer, A. O. (2008). The cusp catastrophe model of psychosis: testing a bifurcation hypothesis of the structure of psychopathology. *J. Abnorm. Psychol.* 117, 814–824.

Crispin, L. M. (2016). School size and student achievement: does one size fit all? *East. Econ. J.* 42, 630–662. doi: 10.1057/eej.2015.2

De Paola, M., Ponzo, M., and Vincenzo, S. (2013). Class size effects on student achievement: Heterogeneity across abilities and fields. *Edu. Econ.* 21, 135–153.

Edgell, J. (1981). Effects of class size upon aptitude and attitude of pre-algebra undergraduate students. *ERIC Document* 203760 Available at: https://eric.ed.gov/?id=ED203760

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., and Willms, J. D. (2001). Class size and student achievement. *Psychol. Sci. Public Interest* 2, 1–30. doi: 10.1111/1529-1006.003

Filges, T., Sonne-Schmidt, , and Viinholt Nielsen, B. (2018). Small class sizes for improving student achievement in primary and secondary schools: A systematic review. *Campbell Systematic Reviews* 14, 1–107. doi: 10.4073/csr.2018.10

Finn, J. D., and Achilles, C. M. (1999). Tennessee's class size study: findings, implications, misconceptions. *Educ. Eval. Policy Anal.* 21, 97–109. doi: 10.3102/01623737021002097

Foreman-Peck, J., and Foreman-Peck, L. (2006). Should schools be smaller? The size-performance relationship for welsh schools. *Econ. Educ. Rev.* 25, 157–171. doi: 10.1016/j.econedurev.2005.01.004

Garrett, Z., Newman, M., Elbourne, D., Bradley, S., Noden, P., Taylor, J., et al. (2004). "Secondary school size: a systematic review" in *Research evidence in education library* (London:EPPI-Centre,SocialScienceResearchUnit,InstituteofEducation,UniversityofLondon)

Gereshenson, S., and Langbein, L. (2015). The effect of primary school size on academic achievement. *Educ. Eval. Policy Anal.* 37, 135S–155S. doi: 10.3102/0162373715576075

Gleason, J. (2012). Using technology-assisted instruction and assessment to reduce the effect of class size on student outcomes in undergraduate mathematics courses. *College Teaching* 60, 87–94.

Grasman, R. P., van der Maas, H. L. J., and Wagenmakers, E. J. (2009). Fitting the cusp catastrophe in R: a cusp package primer. *J. Stat. Softw.* 32, 1–27. doi: 10.18637/jss.v032.i08

Guastello, S. (1984). Cusp and butterfly catastrophe modeling of two opponent process models: drug addiction and work performance. *Behav. Sci.* 29, 258–262. doi: 10.1002/bs.3830290405

Guastello, S. J. (1987). A butterfly catastrophe model of motivation in organization: academic performance. *J. Appl. Psychol.* 72, 165–182. doi: 10.1037/0021-9010.72.1.165

Guastello, S. J. (1992). Clash of the paradigms - a critique of an examination of the polynomial regression technique for evaluating catastrophe -theory hypotheses. *Psychol. Bull.* 111, 375–379. doi: 10.1037/0033-2909.111.2.375

Guastello, S. J. (2001). Nonlinear dynamics in psychology. *Discret. Dyn. Nat. Soc.* 6, 11–29. doi: 10.1155/S1026022601000024

Guastello, S. J. (2002). *Managing emergent phenomena: Non-linear dynamics in work organizations*. Mahwah, NJ: Lawrence.

Guclu, M. (2002). Crowded classroom problem in elementary schools and the solutions to this problem. *Eurasian J. Educ. Res.* 9, 52–58.

Hancock, T. M. (1996). Effects of class size on college student achievement. *College Student Journal* 30, 479–481.

Hattie, J. (2006) The paradox of reducing class size and improved learning outcomes. *International J. Edu. Res.* 43: 387–425.

Hill, M. C. (1998). Class size and student performance in introductory accounting courses: Further evidence. *Issues in Accounting Education* 13, 47–64.

Howley, C. (1996). Compounding disadvantage: the effects of school and district size on student achievement in West Virginia. *J. Res. Rural. Educ.* 12, 25–32.

Howley, C., Strange, M., and Bickel, R. (2000). *When it comes to schooling…Small works: School size, poverty and student achievement*. Knoxville, TN: ERIC Document Reproduction Service.

Hufford, M., Witkiewitz, K., Shields, A., Kodya, S., and Caruso, J. (2003). Relapse as a nonlinear dynamic system: application to patients with alcohol use disorders. *J. Abnorm. Psychol.* 112, 219–227. doi: 10.1037/0021-843x.112.2.219

Jarvis, TJ (2007) *Class Size and Teacher Effects on Student Achievement and Dropout Rates in University-Level Calculus*. Available at: https://www.math.byu.edu/~jarvis/class-size/class-size-preprint.pdf (Accessed December 2014).

Jones, K. R., and Ezeife, A. N. (2011). School size as a factor in the academic achievement of elementary school students. *Psychology* 2, 859–868. doi: 10.4236/psych.2011.28131

Kennedy, P. E., and Siegfried, J. J. (1997). Class size and achievement in introductory economics: evidence from the TUCE III data. *Econ.Edu. Rev.* 16, 385–394.

Kira, I., Barger, B., Shuwiekh, H., Kucharska, J., and Alhuwailah, A. (2019). Cumulative stressors and traumas and suicide: a non-linear cusp dynamic systems model. *Psychology* 10, 1999–2018. doi: 10.4236/psych.2019.1015128

Kira, I., Barger, B., Shuwiekh, H., Kucharska, J., and Alhuwailah, A. (2020). The threshold non-linear model for the effects of cumulative stressors and traumas: a chained cusp catastrophe analysis. *Psychology* 11, 385–403. doi: 10.4236/psych.2020.113025

Koc, N., and Celik, B. (2015). The impact of number of students per teacher on student achievement. *Proced. Soci. Behav. Sci.* 177, 65–70. doi: 10.1016/j.sbspro.2015.02.335

Kokkelenberg, EC, Dillon, M, and Christy, SM (2006) *The effects of class size on student grades at a public university*. Cornell Higher Education Research Institute Working Paper #88. Available at: http://digitalcommons.ilr.cornell.edu/workingpapers/66/.

Konstantopoulos, S., and Shen, T. (2023). Class size and teacher effects on non-cognitive outcomes in grades K-3: a fixed effects analysis of ECLS-K:2011 data. *Large-scale Assess. Educ.* 11, 1–24. doi: 10.1186/s40536-023-00182-8

Krassel, K. F., and Heinesen, E. (2014). Class-size effects in secondary school. *Educ. Econ.* 22, 412–426. doi: 10.1080/09645292.2014.902428

Krueger, A. (1999). Experimental estimates of education production functions. *Q. J. Econ.* 114, 497–532. doi: 10.1162/003355399556052

Kwan, K. (1999). How fair are student ratings in assessing the teaching performance of university teachers? *Assessment and Evaluation in Higher Education* 24, 181–195.

Lammers, W. J., and Murphy, J. J. (2002). A profile of teaching techniques used in the university classroom: A descriptive profile of a US public university. *Active Learning in Higher Education* 33, 54–67.

Lee, V., and Loeb, S. (2000). School size in Chicago elementary schools: effects on teachers' attitudes and students' achievement. *Am. Educ. Res. J.* 37, 3–31. doi: 10.3102/00028312037001003

Lee, V. E., and Smith, J. B. (1997). High school size: which works best and for whom? *Educ. Eval. Policy Anal.* 19, 205–227. doi: 10.3102/01623737019003205

Leithwood, K., and Jantzi, D. (2009). A review of empirical evidence about school size effects: a policy perspective. *Rev. Educ. Res.* 79, 464–490. doi: 10.3102/0034654308326158

Li, G., and Kharabsheh, R. (2022). Dynamic nonlinear system based on complex system theory in the development of vocational education. *Applied Math. Nonlinear Sci.* 8, 961–968. doi: 10.2478/amns.2022.2.0083

Lowenthal, P. R., Nyland, R., Jung, E., Dunlap, J. C., and Kepka, J. (2019). Does class size matter? An exploration into faculty perceptions of teaching high-enrollment online courses. *Am. J. Dist. Educ.* 33, 152–168. doi: 10.1080/08923647.2019.1610262

Lytton, H., and Pyryt, M. C. (1998). Predictors of achievement in basic skills: a Canadian effective schools study. *Can. J. Educ.* 23, 281–301. doi: 10.2307/1585940

Ma, X., and Klinger, D. A. (2000). Hierarchical linear modeling of student and school effects academic achievement. *Can. J. Educ.* 25, 41–55. doi: 10.2307/1585867

Mandel, P., and Sussmuth, B. (2011). Size matters. The relevance and Hicksian surplus of preferred college class size. *Economics of Education Review* 30, 1073–1084.

Martin, J. P. (2015). Moving up in the U.S. News and World Report Rankings. *Change: The Magazine of Higher Learning* 47, 52–61.

Matta, B. N., Guzman, J. M., Stockly, S. K., et al. (2015). Class size effects on student performance in a Hispanic serving institution. *The Review of Black Political Economy* 42, 443–457.

McGuire, K. (1989). School size: the continuing controversy. *Educ. Urban Soc.* 21, 164–174. doi: 10.1177/0013124589021002005

Meier, D. (1996). The big benefits of smallness. *Educ. Leadersh.* 54, 12–15.

Milesi, C., and Gamoran, A. (2006). Effects of Class Size and Instruction on Kindergarten Achievement. *Educational Evaluation and Policy Analysis*, 28, 287–313. doi: 10.3102/01623737028004287

Minnotte, M. C., and Scott, D. W. (1993). The mode tree: a tool for visualization of nonparametric density features. *J. Comput. Graph. Stat.* 2, 51–68. doi: 10.2307/1390955

Molenaar, P. C., and Oppenheimer, L. (1985). Dynamic models of development and mechanistic-organismic controversy. *New Ideas Psychol.* 3, 233–242. doi: 10.1016/0732-118X(85)90017-0

Monks, J., and Schmidt, R. (2011). The impact of class size on outcomes in higher education. *Retrieved from*. Available at: http://digitalcommons.ilr.cornell.edu/workingpapers/114/

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *Futur. Child.* 5, 113–127. doi: 10.2307/1602360

Muir, E. (2000). *Smaller schools: how much more than a fad?* American Educator, 24, 40–46.

Nandrup, A. B. (2016). Do class size effects differ across grades? *Educ. Econ.* 24, 83–95. doi: 10.1080/09645292.2015.1099616

Newman, M., Garrett, Z., Elbourne, D., Bradley, S., Noden, P., Taylor, J., et al. (2006). Does secondary school size make a difference? A systematic review. *Educ. Res. Rev.* 1, 41–60. doi: 10.1016/j.edurev.2006.03.001

Nye, B., Hedges, L. V., and Konstantopoulos, S. (2000). The effects of small classes on academic achievement: the results of the Tennessee class size experiment. *Am. Educ. Res. J.* 37, 123–151. doi: 10.3102/00028312037001123

OECD (2019), *PISA 2018 assessment and analytical framework, PISA*, OECD Publishing, Paris.

Oliva, T., Desarbo, W., Day, D., and Jedidi, K. (1987). GEMCAT: a general multivariate methodology for estimating catastrophe models. *Behav. Sci.* 32, 121–137. doi: 10.1002/bs.3830320205

Olson, J. C., Cooper, S., and Lougheed, T. (2011). Influences of teaching approaches and class size on undergraduate mathematical learning. *PRIMUS* 21, 732–751.

Papageorgiou, G., Stamovlasis, D., and Johnston, P. (2010). Primary teachers' particle ideas and explanations of physical phenomena: the effect of an in-service training course. *Int. J. Sci. Educ.* 32, 629–652. doi: 10.1080/09500690902738016

Poston, T., and Stewart, I. N. (1978). *Catastrophe theory and its applications.* London, UK: Pitman Publishing.

Sapelli, C., and Illanes, G. (2016). Class size and teacher effects in higher education. *Economics of Education Review* 52, 19–28.

Scheerens, J., Hendricks, M., and Luyten, H. (2014). "School Size Effects: Review and Conceptual Analysis" in *School Size Effects Revisited*. (Springer, Cham: Springer Briefs in Education), 7–40.

Schelling, T. C. (1973). Review of structural stability and morphogenesis by René Thom. *J. Polit. Econ.* 81, 1097–1105.

Stamovlasis, D. (2006). The nonlinear dynamical hypothesis in science education problem solving: a catastrophe theory approach. *Nonlin. Dynam. Psychol. Life Sci.* 10, 37–70.

Stamovlasis, D. (2011). Nonlinear dynamics and neo-piagetian theories in problem solving perspectives on a new epistemology and theory development. *Nonlinear Dynamics Psychol. Life Sci.* 15, 145–173.

Stamovlasis, D., Tsaparlis, G., Kamilatos, C., Zarotiadou, E., and Papaoikonomou, D. (2005). Conceptual understanding versus algorithmic problem solving: further evidence from a National Examination. *Chemistry Educ.* 6, 104–118. doi: 10.1039/B2RP90001G

Stewart, I. N., and Peregoy, P. L. (1983). Catastrophe theory modeling in psychology. *Psychol. Bull.* 94, 336–362. doi: 10.1037/0033-2909.94.2.336

Tasker, M. (2003). *Smaller structures in secondary education: A research digest*. Bristol: Human Scale Education.

Thom, R. (1975). *Structural stability and morphogenesis*. Reading, MA: W.A. Benjamin.

Tseng, H. (2010). Has the student performance in managerial economics been affected by the class size of principles of microeconomics. *J. Econom. Econom. Educ. Res.* 11:15.

Tsitsipis, G., Stamovlasis, D., and Papageorgiou, G. (2012). A probabilistic model for students' errors and misconceptions on the structure of matter in relation to three cognitive variables. *Int. J. Sci. Math Educ.* 10, 777–802. doi: 10.1007/s10763-011-9288-x

Van der Maas, H. L. J., Molenaar, P. C. M., and van der Pligt, J. (2003). Sudden transitions in attitudes. *Soci. Meth. Res.* 32, 125–152. doi: 10.1177/0049124103253773

Van der Maas, H. L. J., and Molennar, P. C. (1992). Stagewise cognitive development: an application of catastrophe theory. *Psychol. Rev.* 99, 395–417. doi: 10.1037/0033-295X.99.3.395

Werblow, J., and Duesbery, L. (2009). The impact of high school size on math achievement and dropout rate. *High School J.* 92, 14–23. doi: 10.1353/hsj.0.0022

Westerlund, J. (2008). Class size and student evaluations in Sweden. *Education Economics* 16, 19–28.

Wobmann, L., and West, M. (2006). Class-size effects in school systems around the world: evidence from between-grade variation in TIMSS. *Eur. Econ. Rev.* 50, 695–736. doi: 10.1016/j.euroecorev.2004.11.005

Word, E., Johnston, J., Bain, H. P., and Fulton, B. D. (1990). *Student/teacher achievement ratio (STAR), Tennessee's K-3 class size study: Final summary report 1985–1990*. Nashville, TN: Tennessee State Department of Education.

Yamamori, K., Tokuoka, M., Hagiwara, Y., Oouchi, Y., Nakamoto, K., and Isoda, T. (2021). Effects of class size and provision of learning goals and feedback on students' two-year achievement trajectories. The Japanese. *J. Educ. Psychol.* 69, 297–316. doi: 10.5926/jjep.69.297

Zhang, K. (2016). An exploratory statistical cusp catastrophe model.

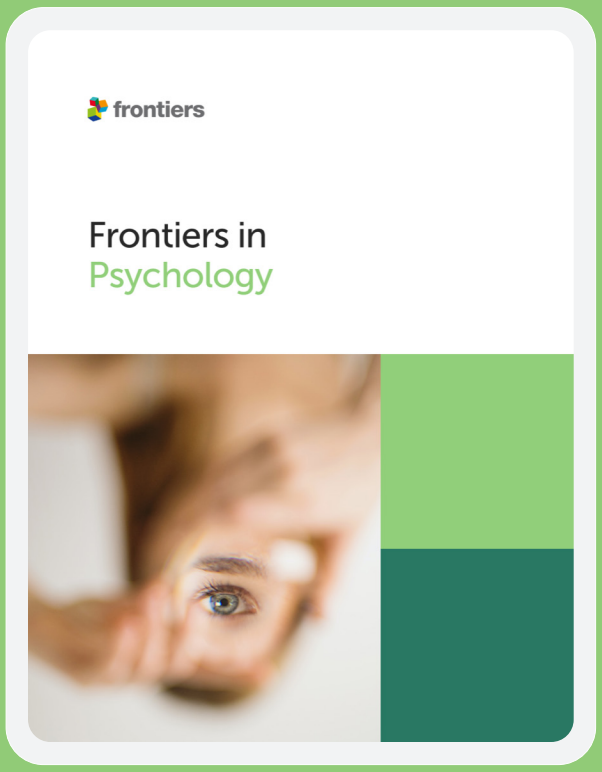
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact