SEMANTIC ALGORITHMS IN THE ASSESSMENT OF ATTITUDES AND PERSONALITY

EDITED BY: Jan Ketil Arnulf, Kai R. Larsen, Oyvind Lund Martinsen and Kim F. Nimon

PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-304-2
DOI 10.3389/978-2-88971-304-2

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

SEMANTIC ALGORITHMS IN THE ASSESSMENT OF ATTITUDES AND PERSONALITY

Topic Editors:
Jan Ketil Arnulf, BI Norwegian Business School, Norway
Kai R. Larsen, University of Colorado Boulder, United States
Oyvind Lund Martinsen, BI Norwegian Business School, Norway
Kim F. Nimon, University of Texas at Tyler, United States

Citation: Arnulf, J. K., Larsen, K. R., Martinsen, O. L., Nimon, K. F., eds. (2021). Semantic Algorithms in the Assessment of Attitudes and Personality. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-304-2

Table of Contents

frontiers in Psychology

EDITORIAL
published: 23 July 2021
doi: 10.3389/fpsyg.2021.720559

Editorial: Semantic Algorithms in the Assessment of Attitudes and Personality

Jan Ketil Arnulf[1]*, *Kai R. Larsen*[2], *Øyvind Lund Martinsen*[1] and *Kim F. Nimon*[3]

[1] BI Norwegian Business School, Oslo, Norway, [2] Leeds School of Business, University of Colorado, Boulder, CO, United States, [3] Soules College of Business, The University of Texas at Tyler, Tyler, TX, United States

Keywords: latent semantic analysis, survey research, organizational behavior, voting behavior, trust, motivation, clinical psychology, artificial intelligence

Editorial on the Research Topic

Semantic Algorithms in the Assessment of Attitudes and Personality

The methodological tools available for psychological and organizational assessment are rapidly advancing through natural language processing (NLP). Computerized analyses of texts are increasingly available as extensions of traditional psychometric approaches. The present Research Topic is recognizing the contributions but also the challenges in publishing such inter-disciplinary research. We therefore sought to provide an open-access avenue for cutting-edge research to introduce and illustrate the various applications of semantics in the assessment of attitudes and personality. The result is a collection of empirical contributions spanning from assessment of psychological states through methodological biases to construct identity detection.

To understand previous research leading up to this issue, one important starting point was the application of machine learning to the assessment of attitudes measured by Larsen et al. (2008). Observing how the output from semantic algorithms could identify high correlations among items, Larsen et al. (2008, p. 3) introduced a mechanism to check for language-driven survey results:

"Manifest validity is expected to support researchers during the data analysis stage in that researchers can compare measures of manifest validity (evaluating the extent of semantic difference between different scales) to item correlations computed from actual responses. In cases where there is little difference between distances proposed by correlation coefficients, the respondents are more likely to have employed shallow processing during questionnaire analysis."

Since then, researchers have expanded the use of semantic similarity of scale items to explore survey responses in a number of ways. Studies have shown that semantics may predict survey responses in organizational behavior (Arnulf et al., 2014, 2018c), leadership (Arnulf and Larsen, 2015, Arnulf et al., 2018b,d), employee engagement (Nimon et al., 2016), technology acceptance (Gefen and Larsen, 2017), and intrinsic motivation (Arnulf et al., 2018a).

In a parallel line of previous research, semantic analysis has been used to complement and extend data from traditional rating scales (e.g., Nicodemus et al., 2014; Bååth et al., 2019; Garcia et al., 2020, Kjell et al., 2019). Since semantic analysis can detect overlap among items and rating scales, they can be used to map relationships and overlap between existing or new scales (e.g., Rosenbusch et al., 2020) and even to detect construct identities and ameliorate the jingle/jangle problem in theory building (e.g., Larsen and Bong, 2016).

While the salient points of several of the articles presented in this Research Topic were semantically similar to prior literature, several others were more diverse (see **Figure 1**).

Edited and reviewed by:
Giovanni Pilato, Institute for High Performance Computing and Networking (ICAR), Italy

***Correspondence:**
Jan Ketil Arnulf
jan.k.arnulf@bi.no

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

Received: *04 June 2021*
Accepted: *28 June 2021*
Published: *23 July 2021*

Citation:
Arnulf JK, Larsen KR, Martinsen ØL and Nimon KF (2021) Editorial: Semantic Algorithms in the Assessment of Attitudes and Personality. Front. Psychol. 12:720559. doi: 10.3389/fpsyg.2021.720559

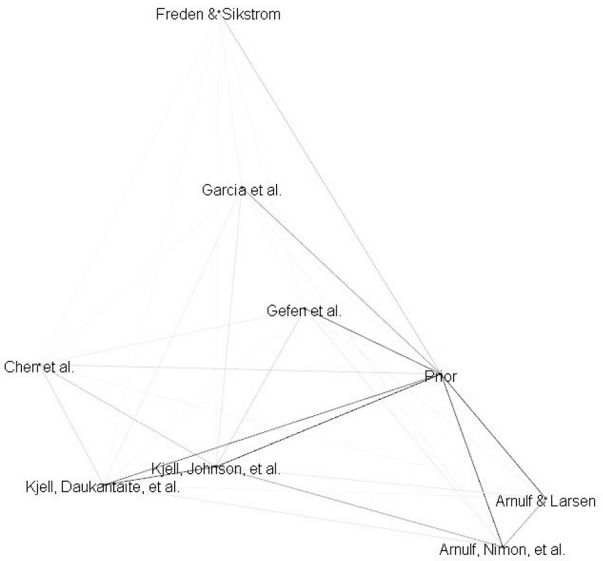


FIGURE 1 | 3D-Plot of Research Topic and prior literature abstracts semantic similarity. Prior encompasses the literature reviewed in the editorial not including the articles contributing to the Research Topic. Darkness of lines represents the magnitude of the cosines resulting from conducting LSA on the abstracts in the Research Topic and prior literature.

Arnulf and Larsen and Arnulf et al. are arguably most similar to the body of literature previously reviewed. In both articles, LSA of survey items predicted survey responses to varying degrees. Arnulf and Larsen's research questioned the capability of traditional survey responses to detect cultural differences. Observed differences in the semantically driven patterns of survey responses from eleven different ethnic samples appeared to be caused by different translations and understanding rather than cultural dependencies. Arnulf et al. similarly found that different score levels in prevalent motivation measures among 18 job types could be explained by differences in semantic patterns between the job types.

Gefen et al. conducted LSA on items sets associated with trust and distrust and found that the resulting distance matrix of the items yielded a covariance-based structural equation model that was consistent with theory.

Kjell O. et al. found that open-ended, computational language assessments of well-being were distinctly related to a theoretically relevant behavioral outcome, whereas data from standard, close-ended numerical rating scales were not. In a similar manner, Kjell K. et al. found that freely generated word responses analyzed with artificial intelligence significantly correlated with individual items connected to the DSM 5 diagnostic criteria of depression and anxiety.

Chen et al. manually annotated Facebook posts to assess social media affect and found that extraverted participants tended to post positive content continuously, more agreeable participants tended to avoid posting negative content, and participants with stronger depression symptoms posted more non-original content.

Garcia et al. applied LSA to Reuter news and Facebook status updates. In the case of the Reuter corpus, the past was devaluated relative to both the present and the future and in the case of the Facebook corpus, the past and present were devaluated against the future. Based on those findings, the authors concluded that people strive to communicate the promotion of a bright future and the prevention of a dark future.

Fredén and Sikstrom applied LSA to voter descriptions of leaders and parties and found that descriptions of leaders predicted vote choice to a similar extent as descriptions of parties.

Nimon provided a dataset of documents from *Taking the Measure of Work* and demonstrated how it could be used to build a LSA space.

As the NLP field continues to develop and mature and the opportunity to automatically transform open-ended data to quantifiable measures, one wonders to what degree the use of rating scales will be warranted in the future. Taken together, the applications demonstrated here go a long way in making free responses accessible to statistical treatment. Similarly, the NLP approaches even seem to allow statistical help in theory building, as the constructs themselves and their relationships with measurement scales may be modeled independently of response data. We invite readers to consider how NLP can advance and/or potentially replace the use of rating scales in the assessment of personality and attitudes.

AUTHOR CONTRIBUTIONS

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

REFERENCES

Arnulf, J. K., and Larsen, K. R. (2015). Overlapping semantics of leadership and heroism: expectations of omnipotence, identification with ideal leaders and disappointment in real managers. *Scand. Psychol.* 2:e3. doi: 10.15714/scandpsychol.2.e3

Arnulf, J. K., Larsen, K. R., and Dysvik, A. (2018c). Measuring semantic components in training and motivation: a methodological introduction to the semantic theory of survey response. *Hum. Resour. Dev. Q.* 30, 17–38. doi: 10.1002/hrdq.21324

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018a). Respondent robotics: simulating responses to likert-scale survey items. *Sage Open* 8, 1–18. doi: 10.1177/2158244018764803

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018b). Semantic algorithms can detect how media language shapes survey responses in organizational behaviour. *PLoS ONE* 13:e0207643. doi: 10.1371/journal.pone.0207643

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Bong, C. H. (2014). Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. *PLoS ONE* 9:e106361. doi: 10.1371/journal.pone.0106361

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Egeland, T. (2018d). The failing measurement of attitudes: how semantic determinants of individual survey

responses come to replace measures of attitude strength. *Behav. Res. Methods* 50, 2345–2365. doi: 10.3758/s13428-017-0999-y

Bååth, R., Sikström, S., Kalnak, N., Hansson, K., and Sahlén, B. (2019). Latent semantic analysis discriminates children with developmental language disorder (DLD) from children with typical language development. *J. Psycholinguist. Res.* 48, 683–697. doi: 10.1007/s10936-018-09625-8

Garcia, D., Rosenberg, P., Nima, A. A., Granjard, A., Cloninger, K. M., and Sikström, S. (2020). Validation of two short personality inventories using self-descriptions in natural language and quantitative semantics test theory. *Front. Psychol.* 11:16. doi: 10.3389/fpsyg.2020.00016

Gefen, D., and Larsen, K. (2017). Controlling for lexical closeness in survey research: a demonstration on the technology acceptance model. *J. Assoc. Inform. Syst.* 18, 727–757.doi: 10.17705/1jais.00469

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikstrom, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Larsen, K., Nevo, D., and Rich, E. (2008). "Exploring the semantic validity of questionnaire scales," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences* (Waikoloa, HI), 1–10.

Larsen, K. R., and Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Q.* 40, 529–551. doi: 10.25300/Misq/2016/40.3.01

Nicodemus, K., Elvevag, B., Foltz, P. W., Rosenstein, M., Diaz-Asper, C., and Weinberger, D. R. (2014). Category fluency, latent semantic analysis and schizophrenia: a candidate gene approach. *Cortex* 55, 182–191. doi: 10.1016/j.cortex.2013.12.004

Nimon, K., Shuck, B., and Zigarmi, D. (2016). Construct overlap between employee engagement and job satisfaction: a function of semantic equivalence? *J. Happ. Stud.* 17, 1149–1171. doi: 10.1007/s10902-015-9636-6

Rosenbusch, H., Wanders, F., and Pit, I. L. (2020). The semantic scale network: an online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychol. Methods* 25, 380–392. doi: 10.1037/met0000244

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 13 September 2018
doi: 10.3389/fpsyg.2018.01623

The Promotion of a Bright Future and the Prevention of a Dark Future: Time Anchored Incitements in News Articles and Facebook's Status Updates

Danilo Garcia[1,2,3], Karl Drejing[3], Clara Amato[1,3], Michal Kosinski[4] and Sverker Sikström[3,5]**

[1] Blekinge Center of Competence, Blekinge County Council, Karlskrona, Sweden, [2] Department of Psychology, University of Gothenburg, Gothenburg, Sweden, [3] Network for Empowerment and Well-Being, Gothenburg, Sweden, [4] Stanford Graduate School of Business, Stanford University, Stanford, CA, United States, [5] Department of Psychology, Lund University, Lund, Sweden

OPEN ACCESS

Edited by:
Tsachi Ein-Dor, Interdisciplinary Center Herzliya, Israel

Reviewed by:
Liliann Manning, Université de Strasbourg, France
Davide Marengo, Università degli Studi di Torino, Italy

***Correspondence:**
Danilo Garcia
danilo.garcia@icloud.com
Sverker Sikström
sverker.sikstrom@psy.lu.se

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

Received: *24 February 2018*
Accepted: *14 August 2018*
Published: *13 September 2018*

Citation:
Garcia D, Drejing K, Amato C, Kosinski M and Sikström S (2018) The Promotion of a Bright Future and the Prevention of a Dark Future: Time Anchored Incitements in News Articles and Facebook's Status Updates. Front. Psychol. 9:1623. doi: 10.3389/fpsyg.2018.01623

Background: Research suggests that humans have the tendency to increase the valence of events when these are imagined to happen in the future, but to decrease the valence when the same events are imagined to happen in the past. This line of research, however, has mostly been conducted by asking participants to value imagined, yet probable, events. Our aim was to re-examine this time-valence asymmetry using real-life data: a Reuter's news and a Facebook status updates corpus.

Method: We organized the Reuter news (120,000,000 words) and the Facebook status updates data (41,056,346 words) into contexts grouped in chronological order (i.e., past, present, and future) using verbs and years as time markers. These contexts were used to estimate the valence of each article and status update, respectively, in relation to the time markers using natural language processing tools (i.e., the Latent Semantic Analysis algorithm).

Results: Our results using verbs, in both text corpus, showed that valence for the future was significantly higher compared to the past (future > past). Similarly, in the Reuter year condition, valence increased approximately linear from 1994 to 1999 for texts written 1996–1997. In the Facebook year condition, the valence of the future was also significantly higher than past valence.

Conclusion: Generally, the analyses of the Reuters data indicated that the past is devaluated relative to both the present and the future, while the analyses of the Facebook data indicated that both the past and the present are devaluated against the future. On this basis, we suggest that people strive to communicate the promotion of a bright future and the prevention of a dark future, which in turn leads to a temporal-valence asymmetrical phenomenon (valence = past < present < future).

Keywords: future, latent semantic analysis, past, present, prevention focus, promotion focus, time-anchored incitements

"I have a dream that one day this nation will rise up and live out the true meaning of its creed: "We hold these truths to be self-evident, that all men are created equal."

I have a dream that one day on the red hills of Georgia, the sons of former slaves and the sons of former slave owners will be able to sit down together at the table of brotherhood.

I have a dream that one day even the state of Mississippi, a state sweltering with the heat of injustice, sweltering with the heat of oppression, will be transformed into an oasis of freedom and justice.

I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character.

I have a dream today!"

Martin Luther King, Jr., 28th of August 1963, at the Lincoln Memorial, Washington, DC, United States

INTRODUCTION

A myriad of theories and empirical studies illuminate our understanding of how we evaluate the past, the present, and the future (e.g., Higgins, 1997; Trope and Liberman, 2000; Caruso et al., 2008; Kurtz, 2008). This line of research has mostly been conducted by asking participants, in experimental conditions, to value imagined positive and/or negative events as occurring either back or forward in time (for some exceptions see: Wilson et al., 2012). At a general level, people usually assign higher values to future events compared to past events (e.g., Trope and Liberman, 2000; D'Argembeau and Van der Linden, 2004; Van Boven and Ashworth, 2007; Caruso, 2010). In the present study, we use data from a Reuter's news corpus and Facebook status updates or "real-life data" (i.e., peoples' actual narratives about past, present, and future events). These real-life data contain statements of both positive and negative events with frequencies that more closely reflect their occurrence in peoples' life, in contrast to controlled experiments with either positive or negative events in equal numbers. We measured the valence[1] of the statements using a semantic statistical method, namely, the Latent Semantic Analysis (LSA; Landauer et al., 1998) algorithm. Thus, the present study makes an important addition to the existing literature because it is based on ecological data from multiple events over time, that we organized in statements of events remembered or imagined to happen in the past, the present, and the future. In sum, we use a larger sample, actual behavior, and data with natural validity that circumvent the limitations of self-reports.

Most experiments on how humans evaluate events in different temporal dimensions ask participants to imagine fictional, yet probable, scenarios. For example, participants are asked to imagine performing a mundane task (e.g., entering data into a computer) and then to rate, at random, the amount of money they would like to get paid if they will perform the task in the future versus if they had already performed the task in the past. Intuitively, one might suspect small differences, however, participants who imagine doing a mundane 5-h task one month in the future demand twice as much more money compared to participants who imagine having completed the same task one month ago (Caruso et al., 2008). This temporal asymmetry is stable across various types of judgments, such as, monetary gain, generosity, and pleasure (e.g., Caruso et al., 2008). In addition, moral transgressions are judged more negatively and deserving more punishment if people imagine them to happen in the future rather than if these transgressions already have happened in the past (Caruso, 2010). In other words, this line of research suggests that when we create a representation of an event happening in the future, both positive and negative events seem to increase in their evaluative magnitude, but to decrease when we imagine that the same events have already happened in the past. One possible reason for this is that people see the future as more exciting and interesting, thus, future events evoke more emotions and curiosity which lead us to make more extreme predictions of the valence of future events (i.e., future heuristic; see Van Boven and Ashworth, 2007; Herbert, 2010). In addition, people in general have a sense of being able to influence the future; therefore, most of us use narratives of the future to promote behavior that is beneficial for ourselves or our group. For example, the Martin Luther King Jr. "I have a dream" speech communicates a positively framed future with desirable values, such as, tolerance and justice. Importantly, the research reviewed here, suggest that the same should hold for negative events, that is, if we are imagining or speaking about a negative event that might happen in the future, we value it more negatively than if we imagine or speak about the same event as if it already have happened in the past (e.g., Caruso, 2010). However, we argue that this temporal asymmetry (i.e., future > past, or past < future, for both positive and negative events) needs to be tested using real life data (cf. Hsee et al., 2014), because in contrast to experimental designs, people typically talk, or write, about different topics and events when making statements about the past, the present, and the future. In other words, the occurrence of positive versus negative past/present/future events in everyday narratives differs from that of experimental controlled designs, which, for good reasons, always present and equal amount of positive and negative events.

These everyday life narratives of past, present, and future events are possible thanks to human beings' unique ability to mental time travel (Suddendorf and Corballis, 1997). These narratives of positive and negative statements of future and past events might influence how humans perceive and recall emotional events. In this context, the ability to react fast to dangerous or negative stimuli is considered essential for an organism to ensure its survival. For example, in a series of experiments (Dijksterhuis and Aarts, 2003), participants detected negatively loaded words more accurately than positive ones, and this was true even when the words were presented subliminally, that is, so fast that the meaning of the words could not be explicitly understood. In other words, suggesting negative valence, rather than positive, as the most common state of being when humans imagine the past and the future. Indeed, a vast amount of research supports the notion that "bad is stronger than good" (Baumeister et al., 2001, p. 323). This includes findings showing that negative emotions, negative feedback, and negative major life events have greater impact in our physical, psychological and social health than positive

[1] Here we use the term high valence as positive and low valence as negative.

ones. This underlying precedence of negativity is also reflected in our language: negative emotions have been shown to be overrepresented in the English language by approximately a 3/5 ratio and this ratio is even stronger (3/4) regarding words describing personality traits (for a review see Baumeister et al., 2001). On this basis, we could expect that an "I have a nightmare" speech would be the most common scenario when people imagine the future.

However, other empirical evidence emphasizes the importance and prevalence of positivity. For example, the analysis of the 5,000 most frequently used words in Twitter, lyrics, books, and the New York Times, suggested an overrepresentation of positive words (Dodds et al., 2011; Kloumann et al., 2012; see also Kramer et al., 2014 for research on emotional contagion in social networks). Moreover, when people imagine a future or past event, positive information is accessed more easily making it more central to the construction of the imagined event (D'Argembeau and Van der Linden, 2004). Perhaps because positive information is more contextual, leading to the construction of more positive and richer imagined future and past events. For instance, despite our tendency to detect negative stimuli faster, negative stimuli are more difficult to remember after longer delays compared to neutral and positive stimuli (Szpunar et al., 2012). That is, showing that humans have a fallacy for a "rose simulated future" (Szpunar, 2010; Szpunar et al., 2012; see also research on self-enhancement and positivity bias; D'Argembeau and Van der Linden, 2004).

This fallacy of a "rosy simulated future," however, might as well be part of what makes people healthy. As the matter of fact, the apprehension of events is also related to peoples' self-regulation (Higgins, 1997; see also Garcia et al., 2010). The "I have a dream" speech is a good example of promotion focused regulation, because it is based on envisioning a successful and bright future (cf. Higgins, 1997). In contrast, people might have a prevention focus when constructing and communicating future events; for example, by envisioning failure and being more vigilant about forthcoming events, in order to avoid or prevent such a dark future (cf. Higgins, 1997). Thus, promotion and prevention focus are important motivators of behavior and even mental health[2] (Higgins, 1997; Amato et al., 2017; Garcia et al., 2017; Amato and Garcia, 2018; see also Walker et al., 2003; D'Argembeau and Van der Linden, 2004). From this perspective, speeches or narratives that envision the promotion of a brighter future or preventing a dark future; both communicate a pleasant or desired state because the individual either envisions a happy and pleasant future or the pleasant relief by avoiding dark or bad outcomes (cf. Higgins, 1997). People, for instance, strive to create legacies that will survive beyond their own existence (Wade-Benzoni and Tost, 2009). Accordingly, having the belief that one has made a difference and will leave the world a better place (cf. promotion focus) leads to the sense of purpose and meaning in life (Wade-Benzoni, 2003; de St Aubin et al., 2004; Grant and Wade-Benzoni, 2009). The motivation to not leave a negative legacy behind (cf. prevention focus) is of equal importance; imposing burdens on powerless others is morally problematic for us humans (Wade-Benzoni and Tost, 2009). Hence, in relation to mental time travel and both positive and negative events, an individual's everyday narratives could be expected to both promote bright futures and prevent dark futures, in turn, devaluating the past.

[2]Even if both types of regulatory focus motivate individuals toward attention to future states (i.e., a brighter future and a less dark future), people's behavior (e.g., action, inaction, counteraction) might differ depending on which type of future is being envisioned.

The Present Study

In summary, findings reviewed here on how we humans evaluate events when we use our ability to mental time travel are complex. First of all, positive events are evaluated as more positive and negative events are evaluated as more negative when these are imagined to happen in the future rather than have happened in the past. Secondly, even if we perceive negative stimuli faster, we selectively prefer to retrieve positive aspects of both past and future events. That being said, since we have a positive heuristic for the future (Herbert, 2010), valence of imagined/constructed future events should be expected to be higher and more positive than recalled/reconstructed past events. Last but not the least, self-regulation theory suggests that both promotion and prevention focus are used to regulate behavior toward desirable positive states (e.g., achieving a desired future or avoiding an undesired future). Hence, narratives and statements from real life, containing a mixture of positive and negative events, could be expected to reiterate a brighter (i.e., promotion focus statements) and less dark (i.e., prevention) future.

The examination of real-life data is important from a methodological perspective (Fischhoff, 1996). For instance, when people reconstruct the past, the present, and the future, the number of positive and negative events is not evenly distributed across temporal dimensions. Since current and predominant views in a society tend to perpetuate themselves through their recurrent presentation in the media (e.g., newspapers, social networks, popular songs) (Garcia and Sikström, 2013a; Garcia et al., 2016), we investigated the temporal valence asymmetry of events using two large text corpora from online newspapers and Facebook status updates by applying the LSA algorithm to quantify the valence of the words (see also Kjell et al., 2018). Specifically, as in previous research we were interested in the valence related to events placed in different temporal dimensions; but in contrast to past research, we did not compare the valence of identical hypothetical events occurring in the past or the future. Instead, we investigated the valence of any events that journalist and Facebook users choose to write about.

MATERIALS AND METHODS

Ethics Statement

This research protocol was approved by the Ethics Committee of Lund University.

Participants

The first data set comprised news stories from Reuters during 1997. We chose this corpus because it was one of the few large news corpus that were public available at the time when the research was conducted. In addition, a few thousand Facebook

users also provided us with 1,183,180 status updates (see the myPersonality project[3]). The Facebook data was collected during 2009 through 2011.

Statistical Method and Procedure

We quantified the valence of temporal markers (i.e., words representing the past, the present, and the future, respectively) using the LSA algorithm. The analyses were conducted in a web-based automated program for analyses of quantitative semantics called semanticexcel,[4] which was developed by one of the authors of this paper. Technical details of how this software generates a semantic representation and predict numbers (valence) from a text based on this representation can be found elsewhere (see Roll et al., 2012, for predicting abstractness; Garcia and Sikström, 2013b, for predicting affectivity scores; and Garcia and Sikström, 2014; Garcia et al., 2015 for predicting personality scores; see also Kjell et al., 2018). Here we just present a brief overview.

Semanticexcel contains semantic representations of several languages, including English. The representation of English used here was generated for the 1997 Reuter news corpus. First, a matrix is generated were rows corresponds to unique single words and each column corresponds to context to the words in the corpus. The rows consisted of the 120,000 most frequency words in corpora, whereas the columns consisted of the contexts of the 10,000 most common words. The contexts of the words were generated from the fifteen words preceding, and fifteen words following, the word in each column. Thus, cells in this matrix represent the frequency of occurrence of a word (rows) within a context of a word (columns). For example, the word "grateful" may have a frequency f_1 in the context "aiding" and a frequency f_2 in the context "accidents." In this way, every word is represented by an array of frequencies of occurrence in each related context to a word.

A basic assumption is that words with similar meaning tend to occur in the same contexts. This implies that the vectors representing similar words should point in similar direction (Sun, 2008). However, to get a good semantic representation this word-by-context sample matrix needs to be compressed to a smaller word-by-semantic dimension matrix, where this smaller matrix tends to create a more generalized semantic representation. We conducted this data compression using Singular Value Decomposition (Strang, 1998), a widespread dimensionality-reduction technique similar to Principal Component Analysis. The resulting matrix is called a semantic space, which describes the semantic relatedness between words. This method has a high level of accuracy, comparable to human performance in different tasks, such as, rating grades (e.g., Landauer and Dumais, 1997, Landauer et al., 1998, Howard and Kahana, 2002). In our analysis, the resulting semantic representation consisted of 120,000 words, where each word is represented in a vector consisting of 100 dimensions.

These representations were used to predict/estimate the valence of each article/status update, respectively, in relation to the time markers (years and verbs were selected as time markers). In the present study, we first identified words related to the past, the present, and the future (i.e., target words). Then we evaluate, using LSA, whether the contexts (the context is defined as the 15 words preceding or following each target word) these words were written in consist of positive or negative words (i.e., the valence). For the sake of clarity, we first briefly describe the rationale behind the chosen time markers, then how we computed valence and then how we did the statistical analyses for testing our hypotheses.

Year-data were divided into categories relative to the publication date. In the Reuter news corpus, the year condition of groups was arranged around the year 1997. By comparing the year that the articles were written, which in the Reuter data was 1997, we identified 1994–1996, as markers of the past, whereas 1998 and 1999, as makers for the future. In the Facebook corpus, this was based on the context content in relation to when the users' status was published. For target words in both Reuter and Facebook data, the verbs were chosen by randomly selecting verbs from McMillan's essential dictionary (Rundell and Fox, 2003). Random selection was used to minimize author bias. This method generated a list of 10 solid past conjugations (see **Table 1**). The English language lack unambiguous usage of the future tense; auxiliary verbs (i.e., verbs that add functional or grammatical meaning and usually accompany a main verb in infinitive) are often needed to imply future tense (Leech, 2004). Some conjugations can be used to describe past, present and/or future (e.g., "Fall" can be used in multiple ways: I Fall [present] and I will Fall [future]). To analyze the future tense, we therefore relied on the fact that this is a modal construction which uses auxiliaries (will or shall) + infinitive (Leech, 2004). Hence, only these two auxiliaries ("will" and "shall") without the infinitive were analyzed to represent the future tense, with the assumption that these are the most frequently used auxiliaries to imply future tense. It should be noted that these auxiliaries can refer to events in the near or far away future, which implies that our data is likely to contain referrals to both near and far away future events. These auxiliaries are shown in **Table 1**.

The method used for predicting the valence of words was multiple-linear regression (y = c*x), where the semantic

TABLE 1 | Verbs and auxiliaries analyzed.

Infinitive	Simple Past (Past)	Past Participle (Present)	Auxiliaries (Future)
Fall	Fell*	Fallen*	Will*
Go	Went*	Gone*	Shall*
Grow	Grew*	Grown*	
Speak	Spoke*	Spoken*	
Be	Was*	Been*	
Write	Wrote*	Written*	
Eat	Ate*	Eaten*	
Drive	Drove*	Driven*	
Do	Did*	Done*	
Choose	Chose*	Chosen*	

* = *Words analyzed in the "Verb" condition.*

[3]https://sites.google.com/michalkosinski.com/mypersonality

[4]www.semanticexcel.com

representations (x) is used as predictors, which are trained on a limited number of words ranked by valence (y). The ANEW (Affective Norms for English Words) wordlist, (Bradley and Lang, 1999) was used to identify one thousand words ranked on valence. Multiple-linear regression was performed between the ANEW list and the semantic space. The resulting regression coefficients (c) can then be used to predict the valence of all words represented in the semantic space. The validity of this method, was estimated with a leave-one-out procedure so that the tested word was removed from the training set, showing a high correlation between predicted and rated scores ($r = 0.62$). Thus, the LSA algorithm generalizes from the evaluation of a small set of ANEW words, to all words in the semantic representation, and thus allows estimation of the valence of a larger number of words, compared to simply counting and affective score based on their ANEW values. We calculated the average valence for words in contexts for target words. This provides a more reliable means of measuring valance; where every single context of a target word has an average predicted valence, rather than the estimated valence of just a target word. In both corpora, 10,000 articles were scanned to obtain the valence of the contexts.

A One-way Analysis of Variance (ANOVA) was computed for each variable in both corpora. In each analysis three conditions were created (Past, Present, and Future). The verbs were assigned into the Past condition if it was written in Simple Past and the Present condition if it was written in Past Participle. Auxiliaries were assigned into the Future condition. Years were assigned to the Past condition if written earlier than the publication date(s), to the Present condition if they were the publication date(s), and the Past condition if written later than the publication date(s). In both corpora, post-hoc two-tailed independent *t*-tests were conducted to examine the difference in valence between the Past and the Present, and the Present and the Future.

RESULTS

Verbs (Reuters)

Ten verbs and two auxiliaries from 10,000 documents produced 14,165 contexts, where some documents produced more than one context. The mean and standard deviation for the valence associated to each group is presented in **Figure 1A**. The frequency

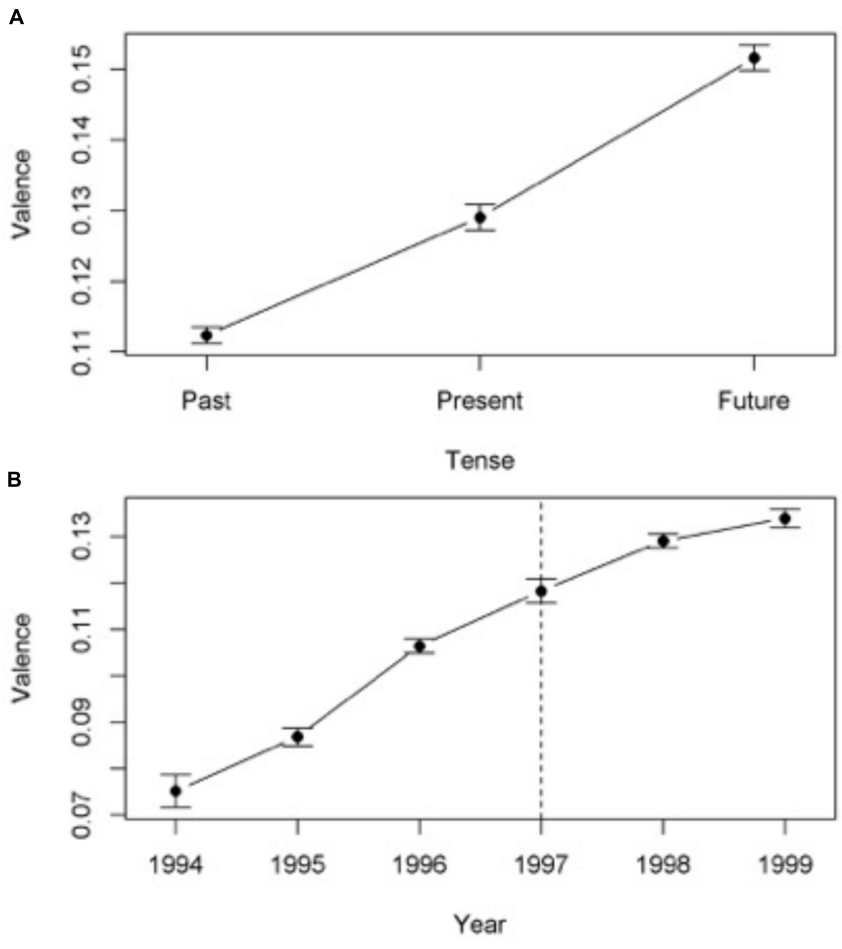


FIGURE 1 | (A) Mean Valence and confidence intervals for Reuter Verbs in Past, Present, and Future tense. **(B)** Mean Valence and confidence intervals for Reuter Years. Dotted line marks the relative publication date.

of occurrence of each verb and condition can be found in **Table 2A**. We conducted an ANOVA to investigate if the valence of the contexts differed between the three conditions: Past, Present and Future [F = 164.0, df = 2, 14162, $p < 0.001$, $\eta^2 = 0.023$, 95% CI(0.018, 0.027)]. Homogeneity of variances was significant at the 0.001 level (*Levene* = 13.17, df = 2, 14162). A two-tailed independent sample *t-test* showed that there was a significant difference between Past and Present [t(11181) = −7.95, $p < 0.001$, d = −0.162, 95% CI(−0.201, −121)] and between Present and Future [t(6504) = −8.98, $p < 0.001$, d = 0.223, 95% CI(−0.272, −0.174)]. In other words, the Present had higher valence than the Past, while the Future had higher valence than the Present. The effects sizes were, however, weak.

Years (Reuters)

Data from six years was analyzed, generating a total of 16,396 contexts.

The mean and standard deviation of the valence for each group can be found in **Figure 1B**. The frequency of occurrence of each year and condition can be found in **Table 2B**. An ANOVA reveled a significant difference in valence between the groups [F = 114.22, df = 5, 16390, $p < 0.001$, $\eta^2 = 0.038$, 90% CI(0.029, 0.039)]. Homogeneity of variances was significant at the 0.001 level (*Levene* = 13.238, df = 5, 16390). A two-tailed independent sample *t-test* showed that there was a significant difference in valence between Past and Present [t(9520) = −7.99, $p < 0.001$, d = −0.221, 95% CI(−0.275, −0.166)] and between Present and Future [t(8438) = −4.55, $p < 0.001$, d = −0.130, 95% CI(−0.182, −0.072)]. In other words, as for the verbs, the Present had higher valence than the Past, while the Future had higher valence than the Present. The effects sizes were, however, weak.

TABLE 2A | Verb frequency, proportions of verbs and proportions of conditions in the Reuter corpus.

Condition	Verb	Frequency	Verb proportions relative to corpus size	Condition proportions relative to corpus size
Future	Shall	992	7.00%	
Future	Will	1990	14.05%	21.05%
Past	Ate	222	1.57%	
Past	Chose	495	3.49%	
Past	Did	930	6.57%	
Past	Drove	446	3.15%	
Past	Fell	1240	8.75%	
Past	Grew	824	5.82%	
Past	Spoke	980	6.92%	
Past	Was	1080	7.62%	
Past	Went	482	3.40%	
Past	Wrote	720	5.08%	52.38%
Present	Been	278	1.96%	
Present	Chosen	743	5.25%	
Present	Done	240	1.69%	
Present	Driven	595	4.20%	
Present	Eaten	169	1.19%	
Present	Fallen	124	0.88%	
Present	Gone	194	1.37%	
Present	Grown	334	2.36%	
Present	Spoken	317	2.24%	
Present	Written	770	5.44%	26.57%
Total		14165	100,00%	100,00%

TABLE 2B | Year frequency, proportions of years and proportions of conditions in the Reuter corpus.

Condition	Year	Frequency	Year proportions relative to corpus size	Condition proportions relative to corpus size
Future	1998	3996	24.37%	
Future	1999	2878	17.55%	41.92%
Past	1996	4156	25.35%	
Past	1995	2861	17.45%	
Past	1994	939	5.73%	48.52%
Present	1997	1566	9.55%	9.55%
Total		16396	100.00%	100.00%

Verbs (Facebook)

Ten verbs and two auxiliaries from 10,000 documents produced 860,127 contexts, where some documents produced more than one context. The mean and standard deviation for the valence associated to each group is presented in **Figure 2A**. The frequency of occurrence of each verb and condition can be found in **Table 3A**. An ANOVA reveled a significant difference in valence between the groups [F = 16717, df = 2, 858668, $p < 0.001$, $\eta^2 = 0.038$, 90% CI(0.037, 0.038)]. Homogeneity of variances was significant at the 0.001 level (*Levene* = 371.55, df = 2,858668). A two-tailed independent sample *t-test* showed that there was a significant difference in valence between Past and Present [t(716660) = 18.98, $p < 0.001$, d = −0.45, 95% CI(−0.050, −0.041)] and between Present and Future [t(458000) = −172.71, $p < 0.001$, d = 0.55, 95% CI(0.546, 0.558)]. In other words, conversely to findings in the Reuters data, the Present had lower valence than the Past. However, in line with Reuters' findings, the Future had higher valence than the Present. The effects sizes were weak or close to moderate.

Years (Facebook)

Data from eleven years were analyzed generating a total of 64,009 contexts. The mean and standard deviation of the valence for each group can be found in **Figure 2B**. The frequency of occurrence of each verb and condition can be found in **Table 3B**. An ANOVA reveled a significant difference between the groups [F = 182.2, df = 10, 63513, $p < 0.001$, $\eta^2 = 0.028$, 90% CI(0.026, 0.030)]. Homogeneity of variances was significant at the 0.001 level (*Levene* = 96.09, df = 2, 63521). A two-tailed independent sample *t-test* showed that there was a significant difference in valence between Past and Present [t(58751) = −27.86, $p < 0.001$, d = 0.48, 95% CI(0.446, 0.513)] and between Present and Future [t(59925) = 29.05, $p < 0.001$, d = 0.10, 95% CI(0.06, 0.146)]. In other words, as for the Reuters' findings, Present had higher valence than the Past, while the Future had higher valence than the Present. The effects sizes were, however, weak.

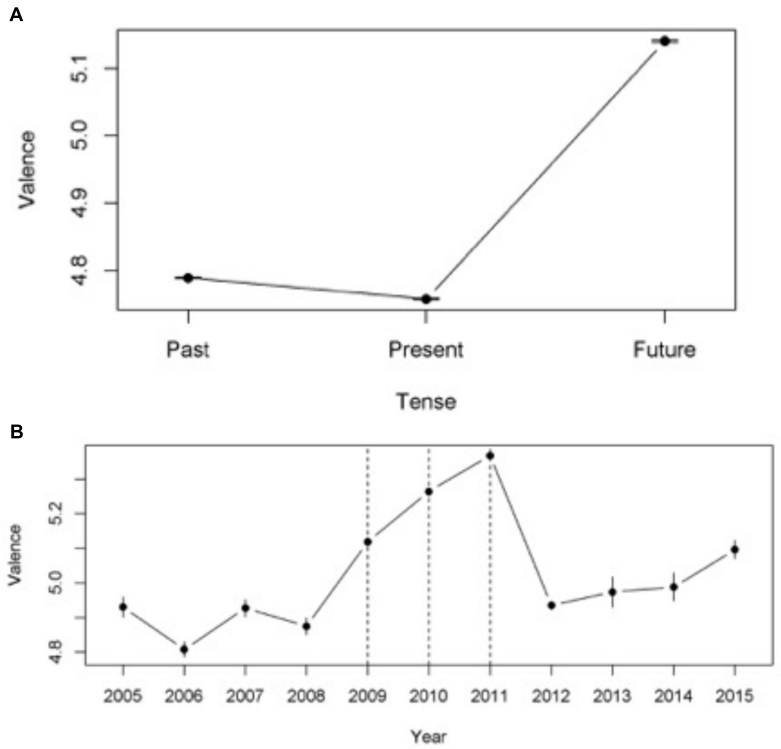


FIGURE 2 | (A) Mean Valence and confidence intervals for Facebook Verbs in Past, Present, and Future tense. **(B)** Mean Valence and confidence intervals for Facebook Years. Dotted line marks the relative publication dates.

DISCUSSION

We investigated the temporal valence asymmetry of events using real-life data (i.e., two large text corpora from online newspapers and Facebook status updates) by applying language processing methods and tools. We identified specific words or target words in the narratives at hand in relations to time markers of the past, the present and the future. We then measured the valence of the contexts (the context is defined as the 15 words preceding or following each target word) in which these target words appeared. Our results using verbs as temporal markers showed, in both the Reuter and Facebook corpus, that valence for the future was significantly higher (i.e., more positive) compared to the past (future > past). Similarly, in the Reuter year condition, valence increased approximately linear from 1994 to 1999. In the Facebook year condition, it is also evident that the valence of the future is significantly higher (i.e., more positive) than past valence. However, for the Facebook data, 2012 did not differ significantly in valence compared to 2007. Nevertheless, the analyses of the Reuters data indicated that the past is devaluated against both the present and the future, while the analyses of the Facebook data indicated that both the past and the present are devaluated against the future. That is, by either devaluating the past against the future or by devaluating the present against the future, both people who engage in the "I have a dream" speech or the "I have a nightmare" speech try always to reach a more pleasant state (cf. Higgins, 1997).

In the present study, the future seems to be valued positively higher than the past, even though current research suggest that evaluations of the future should be more extreme both when it comes to negative and positive events (Caruso et al., 2008; Caruso, 2010). This is even more accentuated in the Reuters data set, which is striking, considering that there was a high likelihood that the sample would include an overrepresentation of lower valence contexts. For instance, news stories that have a more negative valence are twice as likely to be featured in print (Soroka, 2012; see also Trussler and Soroka, 2014). According to the future heuristic, the future is more exciting and interesting, thus, evoking more emotions and curiosity (Herbert, 2010). However, this heuristic only explains that more emotions, both positive and negative, should be associated to texts found in the context of future time-markers. That is, the future heuristic only explains the temporal asymmetry (i.e., past vs. future), not the valence asymmetry found in the present study. Our results, however, might mirror our increased excitement about the future compared to the past (i.e., the future heuristic) in conjunction with our tendency to favor positive information when imagining future events (i.e., positivity bias). This positive excitement about the future is probably based on a solid foundation derived from our concrete perception and physical interaction with the world (i.e., cognitive scaffolding; Herbert, 2010). We humans move forward, and not backward, which in turn might explain why concepts like "progress" and "advancement" are generally associated to something good, while "backward thinking" is

TABLE 3A | Verb frequency, proportions of verbs and proportions of conditions in the Facebook corpus.

Condition	Verb	Frequency	Verb proportions relative to corpus size	Condition proportions relative to corpus size
Future	Shall	42011	4.88%	
Future	Will	100000	11.63%	16.51%
Past	Ate	28653	3.33%	
Past	Chose	6440	0.75%	
Past	Did	100000	11.63%	
Past	Drove	10421	1.21%	
Past	Fell	30519	3.55%	
Past	Grew	6656	0.77%	
Past	Spoke	6423	0.75%	
Past	Was	100000	11.63%	
Past	Went	100000	11.63%	
Past	Wrote	13014	1.51%	46.75%
Present	Been	100000	11.63%	
Present	Chosen	3639	0.42%	
Present	Done	100000	11.63%	
Present	Driven	2500	0.29%	
Present	Eaten	8919	1.04%	
Present	Fallen	8875	1.03%	
Present	Gone	66040	7.68%	
Present	Grown	12203	1.42%	
Present	Spoken	3150	0.37%	
Present	Written	10664	1.24%	36.74%
Total		860127	100.00%	100.00%

TABLE 3B | Year frequency, proportions of years and proportions of conditions in the Facebook corpus.

Condition	Year	Frequency	Year proportions relative to corpus size	Condition proportions relative to corpus size
Future	2012	3753	5.86%	
Future	2013	324	0.51%	
Future	2014	327	0.51%	
Future	2015	398	0.62%	7.50%
Past	2005	645	1.01%	
Past	2006	1016	1.59%	
Past	2007	951	1.49%	
Past	2008	1069	1.67%	5.75%
Present	2009	4759	7.43%	
Present	2010	28515	44.55%	
Present	2011	22252	34.76%	86.75%
Total		64009	100.00%	100.00%

often regarded as bad (see Herbert, 2010, for more examples such as "up vs. down"). Indeed, people seek to make a positive impression upon the world by leaving a legacy that will transcend themselves into future generations (e.g., Wade-Benzoni, 2003; de St Aubin et al., 2004; Grant and Wade-Benzoni, 2009; Wade-Benzoni et al., 2010).

Strengths and Limitations

The quantification of language by extracting words from contexts is a powerful research tool when a large amount of data is available (Landauer and Dumais, 1997; Landauer et al., 1998; Howard and Kahana, 2002; Arvidsson et al., 2011). That being said, research using similar methods in social psychology is limited, making it difficult to compare our findings with previous research. To the best of our knowledge, no previous studies have used the proposed method to examine how people's ability to time travel influences how they evaluate events or rather how it influences the valence in their narratives. One of the strengths of the present study is that we analyzed data from two different domains and found the same overall pattern, that is, that the past is devaluated compared to the future. However, the effect sizes were between weak to moderate. Thus, further experimental and empirical data is needed to confirm or disprove our findings. For instance, it is plausible that narratives of events by non-journalists might give different results. Quoidbach (2013), for example, suggested that there are differences

between the cognitive processes that allow people to look forward and backward in time—imagining new things is generally more difficult than reconstructing old ones from one's personal life. These researchers suggest that, because people find it difficult to imagine themselves changing in the future (e.g., their personality, preferences), people think that it is unlikely the they will actually change (see also Gärling and Gamble, 2012; Garcia et al., 2014). In other words, if people in general find it difficult to change, it is possible that the future is as "rosy" as both the past and the present. In that case, the news and social media data presented here is only a reflection of a contagion of positive emotions for events placed in the future.

Moreover, auxiliary verbs (i.e., verbs that add functional or grammatical meaning and usually accompany a main verb in infinitive) sometimes have other meanings, than implying future tense. For example, "will" or "shall" can in conversational language be used in the present tense to express an ongoing activity that continuous in the near future. Although such exceptions may exist, the most common usage of "will" or "shall" is to describe future events or activities. Common for all verbs, that we used as temporal markers, is also that they are typically used within their denoted tense. Another limitation of the study is that predicting valence using the LSA method may introduce errors in the calculation. Although this is true, we still believe that the LSA is a powerful method that allows automatic measuring of valence with reasonable good accuracy.

Finally, we acknowledge the uneven proportions of extracted verbs and years in the Past, Present and Future conditions. At the most extreme, the years from the Facebook corpus was skewed in the sense that almost 87% of the extracted data was assigned to the Present condition. Most of the data showed the same type of skewness. The verbs from both data sets being the least skewed.

Further Research and Concluding Remarks

Our results open up a number of questions for future research. First, the choice of temporal markers can be further elaborated. Here we chose the time markers based on which words are commonly used as temporal markers in everyday language. Secondly, it would be interesting to replicate the results using different text corpora, such as, literature, novels and short stories, and political speeches. Moreover, there might be cultural differences in how we perceive and represent the past, the present, and the future. For instance, Chinese people seem to recall events from the past in greater detail compared to Canadians (Ji et al., 2009). Also in this line, one's worldview or conception of the world might influence our preference for past or future mental time travel (Ettlin and Hertwig, 2012).

All this being said, our results suggest that the evaluative communication of an event is temporal-valence asymmetrical (that is, valence of an event in time = past < present < future). The outcome, however, depends on whether it can function as incitement for future action or the promotion of behavior (higher valence) or feedback from past actions to avoid or prevent behavior in the future (lower valence): The Time Anchored Incitement Hypothesis (TAIH). We argue that, it might be self-beneficial to the one being the speaker to convey positive evaluative statements about the future that are in line with the legacy she/he envisions to leave for future generations, which in turn also makes the speaker to appear as more appealing and exciting to listeners. After all, we seem to have bias toward a "rosy future." On the other hand, the negative value associated to past events might signal both danger and its proximity (Kyung et al., 2010), thus, focusing attention on improving or even avoiding past behaviors.

AUTHOR CONTRIBUTIONS

DG and CA wrote the paper and revised drafts of the paper. KD and SS conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents, materials, and analysis tools, wrote the paper, prepared figures and/or tables, and reviewed drafts of the paper. MK, KD, SS, and CA reviewed drafts of the paper.

FUNDING

The development of this article was funded by a grant from the Swedish Research Council (Dnr. 2015-01229). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Amato, C., and Garcia, D. (2018). "Regulatory Mode," in *Encyclopedia of Personality and Individual Differences*, eds V. Zeigler-Hill and T. Shackelford (Cham: Springer), 1–9. doi: 10.1007/978-3-319-28099-8_2305-1

Amato, C., Nima, A. A., Mihailovic, M., and Garcia, D. (2017). Modus operandi and affect in sweden: the swedish version of the regulatory mode questionnaire. *PeerJ* 5:e4092. doi: 10.7717/peerj.4092

Arvidsson, D., Werbart, A., and Sikström, S. (2011). Changes in object representations measured by a semantic space method. *Psychother. Res.* 21, 430–446. doi: 10.1080/10503307.2011.577824

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Rev. Gen. Psychol.* 5:323. doi: 10.1037/1089-2680.5.4.323

Bradley, M., and Lang, P. (1999). *Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings*. Gainesville, FL: University of Florida.

Caruso, M. (2010). When the future feels worse than the past: a temporal inconsistency in moral judgment. *J. Exp. Psychol. Gen.* 139, 610–624. doi: 10.1037/a0020757

Caruso, M., Gilbert, T., and Wilson, T. (2008). A wrinkle in time. *Psychol. Sci.* 19, 796–801. doi: 10.1111/j.1467-9280.2008.02159.x

D'Argembeau, A., and Van der Linden, M. (2004). Phenomenal characteristics associated with projecting oneself back into the past and forward into the future: influence of valence and temporal distance. *Cons. Cogn.* 13, 844–858. doi: 10.1016/j.concog.2004.07.007

de St Aubin, E., McAdams, D. P., and Kim, T. (2004). *The Generative Society: Caring for Future Generations*. Washington, DC: American Psychological Association. doi: 10.1037/10622-000

Dijksterhuis, A., and Aarts, H. (2003). On wildebeests and humans: The preferential detection of negative stimuli. *Psychol. Sci.* 14, 14–18. doi: 10.1111/1467-9280.t01-1-01412

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social

network: hedonometrics and twitter. *PLoS One* 6:e26752. doi: 10.1371/journal.pone.0026752 arXiv:1101.5120v5

Landauer, T., and Dumais, S. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295X.104.2.211

Ettlin, F., and Hertwig, R. (2012). Back or to the future? Preferences of time travelers. *Judge. Decis. Mak.* 7, 373–382.

Fischhoff, B. (1996). The real world: what good is it? *Organ. Behav. Hum. Decis. Processes* 65, 232–248. doi: 10.1006/obhd.1996.0024

Garcia, D., Anckarsäter, H., Kjell, O. N. E., Archer, T., Rosenberg, P., Cloninger, C. R., et al. (2015). Agentic, communal, and spiritual traits are related to the semantic representation of written narratives of positive and negative life events. *Psychol. Well Being Theory Res. Pract.* 5, 1–20. doi: 10.1186/s13612-015-0035-x

Garcia, D., Ghiabi, B., Nima, A. A., and Archer, T. (2014). The end of happiness: temporal distance and judgments of life satisfaction in Sweden, Iran, Spain, and El Salvador. *Int. J. Happiness Dev.* 2, 371–382. doi: 10.1504/IJHD.2015.073945

Garcia, D., Kjell, O. N. E., and Sikström, S. (2016). "A Collective Picture of What Makes People Happy: Words Representing Social Relationships, not Money or Material Things, are Recurrent with the Word 'Happiness' in Online Newspapers," in *The Psychology of Social Networking. Identity and Relationships in Online Communities*, Vol. 2, eds G. Riva, B. K. Wiederhold, and P. Cipresso (Berlin: DeGruyter Open).

Garcia, D., Rosenberg, P., Erlandsson, A., and Siddiqui, A. (2010). On lions and adolescents: affective temperaments and the influence of negative stimuli on memory. *J. Happiness Stud.* 11, 477–495. doi: 10.1007/s10902-009-9153-6

Garcia, D., Rosenberg, P., Lindskär, E., Amato, C., and Nima, A. A. (2017). The swedish version of the regulatory mode questionnaire. *Data Brief* 14, 251–254. doi: 10.1016/j.dib.2017.07.050

Garcia, D., and Sikström, S. (2013a). A collective theory of happiness: words related to the word happiness in swedish online newspapers. *Cyberpsychol. Behav. Soc. Netw.* 16, 469–472. doi: 10.1089/cyber.2012.0535

Garcia, D., and Sikström, S. (2013b). Quantifying the semantic representations of adolescents' memories of positive and negative life events. *J. Happiness Stud.* 2012. 14, 1309–1323. doi: 10.1007/s10902-012-9385-8

Garcia, D., and Sikström, S. (2014). The Dark Side of Facebook – dark triad of personality predicts semantic representation of status updates. *Pers. Individ. Diff.* 67, 92–94. doi: 10.1016/j.paid.2013.10.001

Gärling, T., and Gamble, A. (2012). Influences on current mood of eliciting life-satisfaction judgments. *J. Posit. Psychol.* 7, 219–229. doi: 10.1080/17439760.2012.674547

Grant, A. M., and Wade-Benzoni, K. A. (2009). The hot and cool of death awareness at work: mortality cues, aging, and self-protective and prosocial motivations. *Acad. Manag. Rev.* 34, 600–622.

Herbert, W. (2010). *On Second Thought: Outsmarting Your Mind's Hard-Wired Habits*. New York, NY: Crown Publishers.

Higgins, E. T. (1997). Beyond pleasure and pain. *Am. Psychol.* 52, 1280–1300. doi: 10.1037/0003-066X.52.12.1280

Howard, M., and Kahana, M. (2002). A distributed representation of temporal context. *J. Math. Psychol.* 46, 269–299. doi: 10.1006/jmps.2001.1388

Hsee, C. K., Rottenstreich, Y., and Tang, J. (2014). Asymmetries between positives and negatives. *Soc. Personal. Psychol. Compass* 8, 699–707. doi: 10.1111/spc3.12143

Ji, L. J., Guo, T., Zhang, Z., and Messervey, D. (2009). Looking into the past: cultural differences in perception and representation of past information. *J. Pers. Soc. Psychol.* 96, 761–769. doi: 10.1037/a0014498

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2018). Semantic measures: using natural language processing to measure, differentiate and describe psychological constructs. *Psychol. Methods* doi: 10.1037/met0000191 [Epub ahead of print].

Kloumann, I. M., Danforth, C. M., Harris, K. D., Bliss, C. A., and Dodds, P. S. (2012). Positivity of the english language. *PLoS One* 7:e29484. doi: 10.1371/journal.pone.0029484

Kramer, A. D., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. U.S.A.* 7:e29484. doi: 10.1371/journal.pone.0029484

Kurtz, J. (2008). Looking to the future to appreciate the present. the benefits of perceived temporal scarcity. *Psychol. Sci.* 19, 1238–1241. doi: 10.1111/j.1467-9280.2008.02231.x

Kyung, E. J., Menon, G., and Trope, Y. (2010). Reconstruction of things past: why do some memories feel so close and others so far away? *J. Exp. Soc. Psychol.* 46, 217–220. doi: 10.1016/j.jesp.2009.09.003

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Process.* 25, 259–284. doi: 10.1080/01638539809545028

Leech, G. (2004). *Meaning of the English Verb*. Edinburgh: Pearson Education Limited.

Quoidbach, J., Gilbert, D. T., and Wilson, T. D. (2013). The end of history illusion. *Science* 339, 96–98. doi: 10.1126/science.1229294

Roll, M., Mårtensson, F., Sikström, S., Apt, P., Arnling-Bååth, R., and Horne, M. (2012). Atypical associations to abstract words in Broca's aphasia. *Cortex* 48, 1068–1072. doi: 10.1016/j.cortex.2011.11.009

Rundell, M., and Fox, G. (2003). *Macmillans Essential Dictionary for Intermediate Learners*. London: Macmillan Education.

Sikström, S. (n.d.). *LSALAB*. Retrieved 3rd January 2011 from University of Lund. http://www.lucs.lu.se/sverker.sikstrom/LSALAB_intro.html

Soroka, S. N. (2012). The gatekeeping function: distributions of information in media and the real world. *J. Polit.* 74, 514–528. doi: 10.1017/S002238161100171X

Strang, G. (1998). *Introduction to Linear Algebra*, 3rd Edn. Wellesley, MA: Wellesley-Cambridge Press.

Suddendorf, T., and Corballis, M. C. (1997). Mental time travel and the evolution of the human mind. *Genet. Soc. Gen. Psychol. Monogr.* 123, 133–167.

Sun, R. (2008). *The Cambridge Handbook of Computational Psychology*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511816772

Szpunar, K. K. (2010). Episodic future thought: an emerging concept. *Perspect. Psychol. Sci.* 5, 142–162. doi: 10.1177/1745691610362350

Szpunar, K. K., Addis, D. R., and Schacter, D.-L. (2012). Memory for emotional simulations: remembering a rosy future. *Psychol. Sci.* 23, 24–29. doi: 10.1177/0956797611422237

Trope, Y., and Liberman, N. (2000). Temporal construal and time-dependent changes in preference. *J. Pers. Soc. Psychol.* 79, 876–889. doi: 10.1037/0022-3514.79.6.876

Trussler, M., and Soroka, S. (2014). Consumer demand for cynical and negative news frames. *Int. J. Press* 19, 360–379. doi: 10.1177/1940161214524832

Van Boven, L., and Ashworth, L. (2007). Looking forward, looking back: anticipation is more evocative than retrospection. *J. Exp. Psychol.* 136, 289–300. doi: 10.1037/0096-3445.136.2.289

Wade-Benzoni, K. A. (2003). "Intergenerational identification and cooperation in organizations and society," in *Research on Managing Groups and Teams*, Vol. 5, eds M. Neale, E. Mannix, and J. Polzer (Stamford, CT: JAI Press), 257–277.

Wade-Benzoni, K. A., and Tost, L. P. (2009). The egoism and altruism of intergenerational behavior. *Pers. Soc. Psychol. Rev.* 13, 165–193. doi: 10.1177/1088868309339317

Wade-Benzoni, K. A., Sondak, H., and Galinsky, A. D. (2010). Leaving a legacy: intergenerational allocations of benefits and burdens. *Business Ethics Q.* 20, 7–34. doi: 10.5840/beq20102013

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

Wilson, R. E., Gosling, S. D., and Graham, L. T. (2012). A review of facebook research in the social sciences. *Perspect. Psychol. Sci.* 7, 203–220. doi: 10.1177/1745691612442904

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 February 2020
doi: 10.3389/fpsyg.2020.00176

Culture Blind Leadership Research: How Semantically Determined Survey Data May Fail to Detect Cultural Differences

Jan Ketil Arnulf[1]* *and Kai R. Larsen*[2]

[1] *Department of Leadership and Organizational Behavior, BI Norwegian Business School, Oslo, Norway,* [2] *Organizational Leadership and Information Analytics, Leeds School of Business, University of Colorado Boulder, Boulder, CO, United States*

Likert scale surveys are frequently used in cross-cultural studies on leadership. Recent publications using digital text algorithms raise doubt about the source of variation in statistics from such studies to the extent that they are semantically driven. The Semantic Theory of Survey Response (STSR) predicts that in the case of semantically determined answers, the response patterns may also be predictable across languages. The Multifactor Leadership Questionnaire (MLQ) was applied to 11 different ethnic samples in English, Norwegian, German, Urdu and Chinese. Semantic algorithms predicted responses significantly across all conditions, although to varying degree. Comparisons of Norwegian, German, Urdu and Chinese samples in native versus English language versions suggest that observed differences are not culturally dependent but caused by different translations and understanding. The maximum variance attributable to culture was a 5% unique overlap of variation in the two Chinese samples. These findings question the capability of traditional surveys to detect cultural differences. It also indicates that cross-cultural leadership research may risk lack of practical relevance.

Keywords: latent semantic analysis, Likert scales, cross-cultural studies, organizational behavior, semantic versus empirical problems

OPEN ACCESS

Edited by:
Hester Van Herk,
Vrije Universiteit Amsterdam, Netherlands

Reviewed by:
Peter Bevington Smith,
University of Sussex, United Kingdom
Henrik Dobewall,
University of Oulu, Finland
Hiram Calvo,
National Polytechnic Institute, Mexico

***Correspondence:**
Jan Ketil Arnulf
jan.k.arnulf@bi.no

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

Received: *29 July 2019*
Accepted: *24 January 2020*
Published: *18 February 2020*

Citation:
Arnulf JK and Larsen KR (2020) Culture Blind Leadership Research: How Semantically Determined Survey Data May Fail to Detect Cultural Differences. Front. Psychol. 11:176. doi: 10.3389/fpsyg.2020.00176

INTRODUCTION

A simple search for "cross-cultural leadership" through ISI Web of Science returns around 500 hits at the time this is written. An important source of empirical information in these appear to be survey methodology, mostly variations on Likert scale measures. At the same time, a recent methodological development has evolved that sheds a different light on the nature of such data. Relying on digital language algorithms, research on the Semantic Theory of Survey Response (STSR) has opened a way to predict survey patterns *a priori* based on the semantics of the survey items (Arnulf et al., 2014a, 2018a,b; Arnulf and Larsen, 2015; Nimon et al., 2015; Gefen and Larsen, 2017). An unintended but striking finding in one of these studies was that the semantic patterns computed in English were highly predictive also of survey patterns in a Norwegian sample, which raises an important question: If the statistical patterns in survey data are predictable across languages and cultures *a priori*, will such semantically driven surveys detect or neglect cultural differences?

The main tenet of STSR is that responses to survey items will correlate if the items share overlapping meanings. While this has been known and even intended to ensure consistency within

scales, it would lead to contamination and inflated statistics if it happens between scales. Yet this is exactly what previous studies in STSR has found: Using algorithms for text analysis, up to 86% of the variance in relationships between commonly studied variables in leadership research were found to be predictable *a priori* (Arnulf et al., 2014a, 2018b; Nimon et al., 2015).

A peculiar implication of these findings is that if survey response patterns are caused by shared understanding of language, the same patterns should be detectable across languages to the extent that the items are correctly translated. Conversely, if the same survey do not create similar data patterns in samples from different cultures, the differences may be hard to explain even if it would be tempting to assume that differences in data structures are somehow caused by "culture."

This study explores the extent to which cross-national response patterns to a leadership survey are predictable *a priori* through digital semantic algorithms. To achieve this purpose, we have used an instrument that has previously been found to demonstrate semantic predictability, and has also been widely used internationally, the Multifactor Leadership Questionnaire (MLQ) (Avolio et al., 1995; Bass, 1997, 1998). The study will cover native speakers of languages from English through Norwegian, German, Urdu and Chinese, and also compare the responses in native languages to responses in English from parallel respondent groups.

The study serves two purposes: Primarily, it seeks to establish the extent of variation in a cross-cultural leadership survey that can be attributed to semantic relationships. The inverse of this is the maximum amount of variation attributable to cultural factors in a wide sense of the term. Secondarily, this study raises a meta-theoretical question about how cross-cultural differences in leadership can be appropriately captured by our measurement instruments. Understanding the effect of language on leadership across cultures is of great importance in research as well as in practice (Hofstede et al., 2010; Gesteland, 2012; Mendenhall, 2013).

THEORY

The Semantic Theory of Survey Response (STSR) represents a new and hitherto unexplored aspect of survey data (Arnulf et al., 2014a, 2018b; Nimon et al., 2015). Briefly stated, STSR is not about the score levels of items – their purported measurements of latent variables. Instead, the focus of STSR is the semantic structure between the items of measurement instruments. If items in a study – or clusters of items in the form of subscales are semantically related, their mutual score pattern may be influenced by this. Purely semantic patterns in responses have been suggested earlier on theoretical (Feldman and Lynch, 1988; Schwarz, 1999) and experimental grounds (Michell, 1994). With the development of automated algorithms for text analysis, it is now possible to assess the impact and prevalence of this phenomenon in various domains of research (Larsen and Bong, 2016; Gefen and Larsen, 2017; Gefen et al., 2017).

Previous findings in STSR raise a number of methodological and theoretical concerns. What exactly does it imply if the correlation matrix of a survey instrument is predictable *a priori*? It is important here to note that we do not claim that score levels are predictable *per se*. What is predicted are the mutual relationships between the items. Due to the prevalent practice of structural equation modeling in fields like organizational behavior (OB), this means that the input data in the form of correlations or covariance matrix may according to STSR reflect semantic values instead of the purported attitude strength (for an in-depth treatment of this issues, see: Arnulf et al., 2018b,c).

The previous findings in STSR suggested that the factor structures of several instruments were predictable *a priori* due to heavy semantic influences. This is an empirical demonstration of a phenomenon argued conceptually in leadership research. van Knippenberg and Sitkin (2013) argued that the construct of transformational leadership is a tautology, where the dependent variable (leadership effectiveness) is already embedded in the definition and operationalizations of the dependent variables (leadership behaviors). The first study on STSR (Arnulf et al., 2014a) demonstrated empirically that this was in fact the case, and that the problem applied to other measures in leadership and motivation as well.

The meaning of semantic relationships in measurement terms can be understood through the way it works on scale coherence, usually expressed as Cronbach's alpha. Items that share similar meanings (semantic overlap) tend to cluster around similar score levels. In a sense, they are not free to vary because their levels are dependent on each other – a person who believes that today is Friday is not semantically "free" to believe that tomorrow will be a Thursday. The previous studies on STSR found that despite the apparent independence of rotated factors, semantic relationships may still pervade (Arnulf et al., 2014a, 2018b). Measured constructs of leadership and motivation were found to relate semantically, albeit weaker than items within the scales. When this happens, the measured relationships between the latent variables are not free to vary but are mutually "locked." Semantic relationships are not a universal characteristic of all such measurement instruments, as it was not strongly present in a personality inventory. That would imply that respondents to this measurement instrument are less restricted by their previous response in choosing the next response option (Feldman and Lynch, 1988; Maul, 2017; Arnulf et al., 2018c).

The nature and impact of semantic relationships are still not sufficiently understood. So far, we know that survey structures vary between almost complete semantic predictability to almost nothing at all (as in the case of the NEO personality inventory) (Arnulf et al., 2014a). It is likely that the phenomenon is more prevalent where the measures are reflective and the latent variables are social constructions (Arnulf et al., 2018d) than if the measures are formative (Arnulf, 2020). Several studies are going on to determine the variance components most influential in shaping semantic patterns, among others by applying multi-trait-multi-method (MTMM) approaches (Martinsen et al., 2017) but the picture is not yet conclusive.

What seems warranted to claim, however, is that to the extent that statistical patterns are predictable *a priori*, their empirical value is dubious since collecting them does not advance our knowledge (Smedslund, 1988, 2015; Semin, 1989; Elster, 2018).

Semantically determined data patterns reflect agreements across interpretations of items that are common to most speakers. These will be the same across languages if the items in question are translatable.

The focus of STSR, then, is not on the actual score values themselves and the measures that they represent. Instead, STSR is concerned directly with the relationships among the variables – on item level and aggregated between scales.

This is a slightly different perspective from the traditional view on scores as inputs to, e.g., leadership surveys. Here, the score levels are usually collected for at least three purposes: Construct validation, empirical testing of theoretical hypothesized relationships between constructs, and for practitioners, to assess the presence of the theoretical phenomena in a given setting (Nunally and Bernstein, 2007; AERA et al., 2014; Slaney, 2017). For all three purposes, the responses are assumed to be expressions of attitude strength, as originally assumed by Likert (1932). In contrast, STSR is simply concerned with the predictability of semantic overlap between items, as earlier research has demonstrated how information about attitude strength is filtered out when the data structure is semantically determined (Arnulf et al., 2018b).

Culture usually serves as an important context that could presumably modify or even invalidate theoretical claims about leadership (House et al., 2004; Tsui et al., 2007; Mendenhall, 2013; Osland, 2013; Ma and Tsui, 2015). For that reason, the cultural validity of leadership constructs and their relationships to other OB constructs have received extensive attention during recent decades. There have also been a number of discussions about the methodological opportunities and pitfalls imminent in such research (House et al., 2004; Kirkman et al., 2006; Mansour et al., 2006; Hofstede et al., 2010). The present study does not aim at a comprehensive review of previously discussed opportunities and pitfalls. The focus here is on a specific problem with possibly wider ramifications: That cross-cultural research on OB may be trapped in semantic tautologies that obstruct real empirical insights.

Semantically Determined Relationships

The Semantic Theory of Survey Response posits that the most obvious reason for correlations between survey items will be that they overlap in meaning (Arnulf et al., 2015). If a person thinks that today is Thursday, the person is also likely to think that tomorrow is Friday. This is not an empirical, but a semantic relationship – the one follows from the other (Semin, 1989; Smedslund, 1994; McEachrane, 2009). Ideas about weekdays may be blatantly obvious, but fuzzier examples of weaker relationships exist. People who say that they enjoy their jobs will also be less likely to look for new jobs – to want to keep a job is part of the meaning of liking one's job. Since some people still look for other opportunities even while liking their present jobs, there will not be a perfect correlation between the two. These are examples of semantic relationships with various strengths.

"Semantics" is the branch of linguistics and logic concerned with meaning (Semin, 1989; Deerwester et al., 1990). The term "semantic relationship" usually implies one of two related meanings: Either the lexical definition of words and terms, as when using a dictionary, or the logical implication of one term from another as when explaining an argument. Until recently, semantics has been a domain for linguists and logicians. With the development of digital techniques for natural language processing, semantics has also become an important part of information technology (Landauer et al., 1998; Landauer, 2007; Dennis et al., 2013; Zhang et al., 2013). There now exist a variety of algorithms that can be used to index and compare the meaning of texts. Most readers are familiar with them in applications such as internet search engines. They can also be used for a number of advanced purposes such as automated translations or to establish ontologies – automated taxonomies that classify and organize knowledge about domains of discourse. Digital text algorithms can be used as tools to analyze and compare texts (Larsen and Bong, 2016; Gefen et al., 2017). They are relatively impartial in the sense that they follow transparent rules that will yield the same results across texts if applied in identical ways.

Using digital algorithms for text analysis, previous studies have found that widely used constructs within the OB domain are in fact semantically determined (Arnulf et al., 2014a, 2018a; Arnulf and Larsen, 2015; Nimon et al., 2015; Kjell et al., 2019). Digital algorithms take texts as their input and can perform computations on their meanings, comparing and grouping text according to quantitative measures of similarity. Digital algorithms have demonstrated the semantic link between constructs such as transformational leadership, LMX, 2-factor leadership, intrinsic motivation, OCB, and commitment (Arnulf et al., 2014a, 2018a). The specific semantic algorithms used in this study are further explained in the methods section.

The problematic side of semantic relationships is that they are basically only parallel or re-iterated versions of the same underlying propositions. This is easiest to see in the example concerning weekdays. If we know someone's belief about which day we have today, we can predict all other statements that place the other 6 weekdays. It is also worth noticing that this is not limited to one language. The same sentences will be true in any other language as long as the language has words making up a 7-day week. That is because the propositional structure of the sentence is on a more abstract level than the words themselves. As long as the propositional structure is kept intact, the actual wording does not matter, whether within nor between languages.

While the example about the weekdays may be easy to understand, it gets harder when propositions only share some, but not all of their meaning. This is, however, the most likely reason for even weak correlations between survey items. If a respondent describes satisfaction with her job, the actual meaning of this is, among other criteria, that this job is preferable to other jobs. Hence, there is every reason to assume that job satisfaction will be negatively correlated with the intention to switch jobs. The correlation may however be far from perfect because "preferring this job to other jobs" is only one of many explanations for job satisfaction.

To the extent that survey data represent semantic relationships instead of attitude strength, they will not easily detect cultural differences. Most semantic relationships are translatable across modern languages and certainly in the field of organizations and leadership. To the extent that semantically determined

correlations and other data structures are replicable across cultures and languages, it may only tell us that the semantic structure of the survey was correctly reproduced across these languages.

Therefore, Hypothesis 1:
Correlations in leadership surveys that are semantically determined in one language will be semantically predictable to a significant degree across all national languages and working environments.

Cultural Differences in Survey Data

Conversely, if structures in survey data can be supposed to convey culturally determined patterns, they need to display variation that is unique to the linguistic or ethnic group as different from other, culturally unrelated samples (House et al., 2004). A simple version of this argument is frequently implied in the analysis of cross-cultural samples, in that differences between populations with different ethnic or other demographic characteristics are taken as indications of cultural similarities or differences.

A previous study has shown that while a range of respondent properties may influence score levels on leadership surveys, the ensuing correlation matrix has a tendency to converge around a structure predicted by semantics (Arnulf et al., 2018b). Our focus here is solely on the degree to which nationalities and languages influence the degree to which semantics can explain the item correlation matrix.

Languages pose a complex methodological challenge in research on management and OB (Harzing et al., 2011; Zander et al., 2011). The initial concern was to preserve the meaning of items when surveys were translated. Hence, it was suggested that surveys should be translated and independently translated back to assure that the meaning of the original items were preserved (Brislin, 1970; Herdman et al., 1997). More advanced developments in this field have recognized the insufficiency of this approach (Behr et al., 2016). While translation-back-translation may even create problems instead of solving them, a bigger problem arises when there is no accurate expressions in the second language for the target item of the original survey. For example, key modern-day English terms from the workplace do not necessarily exist or have the same meaning in other languages. The word "leadership" does not exist in, e.g., French, Italian or Japanese, but are usually substituted with the English word. The German counterpart for leadership ("Führung") was politically contaminated and has largely been replaced with the English word "Management" (Arnulf et al., 2018d), but with slightly different meanings – what the linguists call "false friends" (Enfield, 2007).

While most survey items do not use such high-level concepts, they may still require the import of new linguistic constructions or professional expressions with limited public accept into the second language. In such cases, the survey may actually be translatable on one level and still difficult to understand at other levels (Behr et al., 2016). Differences in response statistics due to problems in understanding and translatability may appear as "cultural differences" but simply signal lack of understanding by the respondents.

Thus, Hypothesis 2:
Differences in survey response statistics between different ethnic and linguistic groups can be empirically explained by lack of understanding of the item texts, rather than systematic cultural differences.

Idiomatic Equivalence

While items may be accurately translated on a surface level, proper translations need to address the underlying propositional structure (Hanks, 1996; Behr et al., 2016). For example, a proverbial expression such as "to judge a book by its cover" is not actually about books, and is at the surface level easy to translate into any language that includes the concepts of 'judgement' and 'books.' If the underlying metaphorical phrase does not exist in the focal language or is less frequently used, respondents are less likely to fill out a survey appropriately. For example, translating the idiom to a language like Norwegian, will yield "å dømme en bok etter omslaget." Many Norwegians will actually know of the English idiom, but a search for the phrase at Google.no will yield articles literally about whether consumers buy books based on the attractiveness of the cover. The requirement of idiomatic equivalence is common knowledge to most translators but it bears special relevance to the problem of semantic determination of survey response statistics (Arnulf et al., 2018d). If the translation departs from the idiographic essence, it can be inaccurate even when the superficial words look similar. In such cases, different statistics will not signal cultural differences but inaccurate translation.

The problem of idiomatic equivalence is therefore a core issue in cross-cultural leadership. Are different ways of conceptualizing work place phenomena simply different expressions of the same underlying theoretical "constructs," or do they actually imply different cultural constructions of the work place? Only the latter case would indicate a true cultural difference, but it will be harder to detect within the conditions of the survey items itself. In this sense, survey data are "thin" in the sense of Geertz (1973) – they do not carry information about whether they are methodological artifacts or indicative of true cultural differences.

The Language Relativity Hypothesis

The proposition that native languages construct the experience in unique ways has had a long history in the humanistic and social sciences (Gumperz and Levinson, 1996). Most frequently attributed to Whorf (1956), there have been recurrent controversies about this topic (Lucy, 1996). The most extreme version of this hypothesis asserts that we do not experience what we have no words for, and conversely have richer experiences where we have more nuanced words. While this extreme version is probably not true (and also not endorsed by many), an increasing volume of empirical research seems to document that native languages do influence our cognitive functions and verbal interactions (Slobin, 1996; Boroditsky, 2011; Sidnell and Enfield, 2012; Gentner, 2016). A modified version of the linguistic relativity hypothesis seems to be documented and allow at least two important predictions: The first is that different languages provide different tools for perception and experiences. Language structures do not in themselves open or block experience, but

they do guide attention and emphasis in culturally determined ways (Slobin, 1996). Languages are culturally accumulated tools and may be one of the most important sources of acculturation (Lakoff, 1987; Cavalli-Sforza, 2001; Pinker, 2008). While foreign language constructions may be expressible to some degree in every other language, the attention, nuances and importance of verbal content may be determined by one's native language. Secondly, cognition and behavior in bilingual humans is influenced by the language in which they use in interactions (Hanks, 1996; Arnulf et al., 2014b). It follows from this that the most truly "culturally" determined responses detected in survey statistics are likely to be elicited from respondents to surveys in their native languages (Boroditsky, 2001, 2011; Boroditsky and Gaby, 2010; Fausey et al., 2010; Costa et al., 2017). Survey designs that use common corporate languages (usually English) may omit the translation problem, but will risk missing the truly "cultural" identity of a bilingual respondent. One way to ensure that differences in survey responses are truly culturally determined would be to combine two approaches, a native language and a corporate language approach. If the two conditions yield response patterns that are unique to the ethnic group, one may safely assume that it taps native language understanding while at the same time adheres to the same item structure that is presented to all participants (original language).

From the point of view of STSR, this sets up two criteria for determining cultural uniqueness in response patterns. First, the response pattern of the target group (e.g., Chinese) needs to be significantly less predictable by the language used in the algorithms (e.g., English). Second, there needs to be an identifiable shared proportion of variance between the target group surveyed in its native language (e.g., Chinese) and in the language used by the algorithm (e.g., English).

> Thus Hypothesis 3:
> *Samples of respondents who do not have English as their native language will display unique common variance that is neither explained by semantic algorithms nor by response patterns from unrelated cultures.*

In what follows, we will test the three hypotheses by applying text algorithms to a frequently used measurement instrument in leadership research and compare its predictive capabilities across a panel of diverse languages and ethnic groups.

MATERIALS AND METHODS

Measures

Survey Instrument

The survey used for this study was the Multidimensional Leadership Questionnaire (MLQ) commonly used in research on transformational leadership (Tejeda et al., 2001; Piccolo et al., 2012). This instrument was used for two main reasons: For one, it has previously been shown to be semantically determined to a substantial degree (Arnulf et al., 2014a, 2018c). Secondly, it exists in a series of authorized non-English versions, frequently used in cross-cultural research and as basis for claims about cross-cultural validity of its main constructs[1].

The MLQ was administered as a web-based survey, all items on a 5-point Likert scale and every item was fully labeled.

Semantic Algorithms

Following previous studies in STSR, we used two main types of algorithms. One is a corpus-based approach often termed MI (Mihalcea et al., 2006), the other is a vector-based approach called Latent Semantic Analysis (LSA) (Deerwester et al., 1990). These algorithms are extensively published and described methodologically elsewhere in articles on semantics in psychometrics (Arnulf et al., 2014a, 2018a; Larsen and Bong, 2016; Gefen and Larsen, 2017; Gefen et al., 2017), but their main features are presented briefly here.

The MI algorithm (Mihalcea et al., 2006) extracts meaning from a lexical database called WordNet (Poli et al., 2010). It parses sentences into words and detects part-of-speech to better detect the correct category for the words in WordNet. Word specificity refers to the specific meaning of words (e.g., collie and sheepdog) versus generic concept words (e.g., animal and mammal). Specific words are given higher weight than abstract concepts (such as animal). The British National Corpus (Sparck-Jones, 1972) is used to calculate inverse document frequency (Sparck Jones, 1986). The version of the MI algorithm used here is the same as that used in Larsen and Bong (2016), which along with path similarity averages word-similarity metrics from Wu and Palmer (1994), Jiang and Conrath (1997), and Lin (1998). These metrics were created to measure word relatedness and similarity by calculating the shortest distance between given words' synsets (sets of synonymous words) in the WordNet hierarchy; the shorter the distance between words, the higher the similarity score. For implementation details on the MI algorithm, please see Larsen and Bong (2016).

Through a combined calculation of lexical distances and the syntactic structure of the sentences, the MI algorithm will assign a number signifying overlap in meaning between any two survey items (Mihalcea et al., 2006). This number will always be between 0 and 1.00, where a higher number indicates greater overlap of meaning. The numbers are structurally similar to correlations but cannot take negative values and are also different from correlations in that they do not depend on co-variation– they are strict assessment of the overlap of meaning.

The LSA algorithm does not make any use of pre-defined lexical information. Instead, it "extracts" meaning from large samples of existing text called "semantic spaces" (Dennis et al., 2013; Gefen et al., 2017). These semantic spaces are made up of hundreds of millions of words that have been collected from a defined text universe, such as newspaper articles, textbooks or scientific publications. These text samples are turned into a word-by-document matrix, then further reduced in a statistical technique called "singular value decomposition" (SVD). The similarity of texts such as survey items can then be determined by projecting the items texts onto the SVD-transformed matrices (Gefen et al., 2017). The output from LSA are the cosines of the

[1] http://www.mindgarden.com/

compared items in these matrices. Like the MI values, the LSA values usually fall in the range between 0 and 1.00 even though they occasionally do take negative values. These negative values are however not the same as negations.

All these algorithms are still inferior to humans in their ability to detect meaning (Landauer, 2007). Since the LSA output is dependent on the semantic space applied, we usually compute LSA values from multiple semantic spaces to approximate the understanding of human speakers. Finally, by combining MI and LSA values in multiple regression, we can approximate the semantic understanding of human subjects as a combination of lexical and domain-specific knowledge, as shown by previous authors (Arnulf et al., 2014a, 2018a). As will be discussed below, the semantic algorithms are still inferior to language parsing in humans. While the data sources (WordNet and newspaper articles) used in the algorithms are not unbiased (see, for example Baeza-Yates, 2018), none of these sources were designed or collected with knowledge that they would one day be used to evaluate survey items.

Despite their shortcomings, the algorithms pose a sort of "impartial" standard for semantic structures in that they are transparent and completely rule-based, leaving out subjective measurement errors (Stark, 2018).

Human Respondent Samples

Because of the cross-cultural, multi-language nature of this study, we aimed to obtain a broad and still balanced set of sub-samples. The semantic algorithms were all computed in English and the prevalently used leadership survey MLQ was also originally published in English. Hence, we chose English as the basic language of the analysis. This is also in line with a prevalent practice of using English as corporate language across the world (Harzing et al., 2011; Zander et al., 2011).

We sought to compare groups with native languages of differing distance to English, ranging from proximal to distant in terms of language families. We obtained one sample of 146 native speakers of English to represent the baseline computed by the algorithms. The samples with native languages closest to English were obtained in Norwegians (N = 1,226 sampled in Norwegian and 180 Norwegians responding in English) and Germans in German (N = 59, none in English). These languages share the Indo-European language roots of English and are assumed to be distinct but close (Cavalli-Sforza, 2001). As a more remotely related sample, we chose Pakistanis responding in Urdu (N = 111) and Pakistanis responding in English (N = 108). Urdu is another Indo-European language but with much more distant relationship to English than the other two (Cavalli-Sforza, 2001). Finally, we chose Chinese (N = 259 Chinese responding in Mandarin and 240 Chinese responding in English) as the sample with the greatest linguistic and cultural distance from English (Needham and Harbsmeier, 1998; Cavalli-Sforza, 2001; Norenzayan et al., 2002). Through the data sampling procedure (see below) we also had three other mixed sub-samples: 45 other Europeans responding in English, 49 Indian nationals in English (who stated other options as their native language, e.g., Tamil, Malayalam, etc.), and 58 non-Chinese East Asian citizens responding in English (mostly Indonesians, Koreans, and Japanese).

The data mainly stem from leadership surveys carried out in four globally present companies. The employees from these companies were mainly staff working with banking, engineering, sales and administrative functions such as accounting and HR. The responses were mostly sampled from locations in Norway, Dubai, India, Singapore, Korea and China. To balance the design, there were three convenience samples: The native speakers of German and about a third of the native speakers of English were recruited through the network of the researchers. The native English speakers were a mixed group of people from the United States and the United Kingdom, with a small number of Indian and Singaporean citizens who described their native languages as "English." Half of the Pakistani respondents using Urdu were working at an engineering college in Pakistan, but another half were first generation immigrants in Norway working in diverse professions.

For the whole sample, the mode of the age group was 35–44 years, and 58% were male. While 68.1% described themselves as non-managers, 25.1% were middle managers, 4.1 were upper management and 2.7% described themselves as executive level.

Analytical Strategy

As previously stated, our analysis aims at exploring the degree to which the observed item response matrices (the dependent variables) of our various samples are explained in regression equations using the semantic indices as independent variables.

RESULTS

We first established the characteristics of each sample in terms of demographics, linguistic background and the main score levels on the leadership scales of the MLQ. **Table 1** presents these values in overview.

An ANOVA analysis shows that the differences in score levels between the samples are statistically significant, but not large. For all samples, the transformational leadership score averages are in the range of 3.3 – 3.7. The score levels of transformational leadership are universally higher than the sample scores for transactional leadership, where the range is wider (2.6 – 3.5). The range of Laissez-faire is 1.7 – 3.2, and the outcome scores range between 3.5 and 4.1. More importantly, the differences in means appear to be random variation without any systematic relation to sample size or cultural distance from native speakers of English.

The full version of the MLQ contains 45 items (Avolio et al., 1995). This turns into a matrix of (45*44)/2 = 990 unique item correlations. The semantic method addresses these relationships, which are also important to most prevalent statistical models. The correlations or co-variances between items and scales are commonly used to build statistical models in survey research (Jöreskog, 1993; Borsboom, 2008; MacKenzie et al., 2011; Lamiell, 2013; Slaney, 2017; Van Dierendonck et al., 2017). To the extent that these are semantically determined, the semantic influence will be retained in all subsequent models.

TABLE 1 | Sample characteristics and score levels.

			Mean leadership score levels			
Experimental group	***N***	**Male/Female**	**Transformational leadership**	**Transactional leadership**	**Laissez-faire**	**Outcome scores**
English native speakers	146	70%/30%	3.4	2.9	2.0	3.5
Norwegians in Norwegian	1226	51%/49%	3.7	3.0	2.1	3.6
Norwegians in English	180	82%/18%	3.5	2.6	1.7	3.7
Germans in German	59	61%/39%	3.3	3.1	2.3	3.5
Other Europeans in English	45	80%/20%	3.6	2.9	1.9	3.6
Pakistanis in Urdu	111	n/a	3.7	3.5	3.2	3.8
Pakistanis in English	108	n/a	3.7	2.8	1.9	4.1
Indian nationals in English	49	82%/18%	3.4	2.9	1.9	3.5
Chinese in Chinese	235	57%/43%	3.5	3.0	2.0	3.5
Chinese in English	240	61%/39%	3.6	3.0	1.7	3.7
East Asians in English	58	76%/24%	3.6	3.0	1.9	3.6
Total dataset	**2513**	**58%/42%**	**3.6**	**3.0**	**2.1**	**3.6**

TABLE 2 | Predicted variation of the correlation matrix for each linguistic sub-sample, compared with a principal component analysis (PCA) of each sample.

Experimental group	**Predicted in linear regression (adj R^2)**	**Predicted in GLM (adj R^2)**	**Predicted in GLM full factorial (adj R^2)**	**PCA Eigenvalues > 1**	**Variance explained by the PCA factors**	**PCA Visual Scree factors**
English native speakers	0.84	0.87	0.91	7	70	1
Norwegians in Norwegian	0.79	0.86	0.91	6	59	1
Norwegians in English	0.66	0.77	0.89	11	71	1
Germans in German	0.67	0.73	0.80	9	75	3
Other Europeans in English	0.77	0.83	0.94	8	82	3
Pakistanis in Urdu	0.11	0.21	0.31	12	72	5
Pakistanis in English	0.43	0.55	0.71	11	76	3
Indian nationals in English	0.73	0.78	0.83	8	78	1
Chinese in Chinese	0.54	0.59	0.67	10	69	2
Chinese in English	0.72	0.77	0.86	10	67	3
East Asians in English	0.55	0.67	0.74	10	85	2
Total dataset	0.79	0.85	0.92	6	57	3

We therefore regressed the semantic values on the item correlation matrix for each sample. This can be done in three ways (Arnulf et al., 2018a): The first is a multiple linear regression where we use all the semantic information but in a purely linear model. This approach probably underestimates the semantic influence, because the semantic algorithms available at present cannot take context into consideration. Human speakers use context as an important signal to differentiate between different meanings of the same words. To emulate this, we may set up a general linear model (GLM) that allows the equation to "know" which scale any item belongs to. This comes close to human contextual understanding and is justified because the scale belongingness is significantly predictable by the algorithms (Arnulf et al., 2014a). We try two types of GLM: In the first model, we only use the main effects on the variables but set the constants as fixed within the scales. In the second model we use the full interactions between the variables. The final approach obviously risks overfitting the model. We therefore report the results of all three models, taking the linear model as a lower and the GLM estimates as an upper limit to the "true" effect of semantics on the correlation matrix.

Hypothesis 1 stated that "Correlations in leadership surveys that are semantically determined in one language will be semantically predictable to a significant degree across all national languages and working environments." This is tested and listed for each of the language subgroups in **Table 2**.

All regression models are significant ($p < 0.001$), and therefore support hypothesis 1. However, there are differences that could conceivably be due to culture. The same semantic values predict the different linguistic groups in a range from 84% in the case of native English speakers down to 11% for Pakistanis responding to a version in Urdu. In fact, there is a strong negative relationship between semantic predictability and the complexity of the factor structure when the samples are subjected to a principal component analysis (PCA): The more semantically predictable the dataset appears to be, the lower the number of Eigenvalues above 1 and the lower the number of factors visually identifiable in the Scree plots.

While this could indicate different cultural backgrounds in leadership cultures, the more parsimonious interpretation is that it could be noise due to lack of understanding. There are particularly three conspicuous facts that point in this direction:

The Norwegians are strongly semantically predictable, but more predictable in their native language Norwegian than in English. The Pakistanis seem only vaguely compliant with the semantics when answering in Urdu, but much more so for those who are allowed to answer in English. The two Chinese samples, that linguistically and culturally should be more distant from the Anglo-Saxon culture than the Pakistanis, are much more influenced by semantics and also here, those surveyed in English seem more semantically predictable than those responding in Chinese. Moreover, the Indian nationals, who arguably are not culturally very distant from the Pakistanis, are very semantically predictable when responding in English. In short, there does not seem to be a systematic pattern that explains how samples depart from the semantically expected.

Using the R^2's tells only part of the story. If the departure from semantically expected correlations are due to noise, the residuals will be fairly random, and the systematic part of the variation will still be semantics. The first way to test this is to see how well the semantically predicted correlations actually match the real survey correlations. Central to leadership research is an interest in the mutual impact of leadership behaviors on purported outcomes (March and Sutton, 1997; Dumdum et al., 2002; Hansen et al., 2013; Arnulf et al., 2018d). Since the MLQ contains a separate scale for outcomes, we can average the correlations between each leadership behavior and the outcome measures and compare these to the values predicted in the respective regression models. We can thereby estimate how the semantic values predict the theoretically proposed relationships between leadership behaviors and outcomes in the employees. This is displayed in **Table 3**.

The overview shows that the correlations between the various leadership behaviors and the outcome values are almost equally well predicted across the linguistic sub-samples, ranging from almost identical in the case of GLM to somewhat less precise in linear regression. The finding is in accordance with the theoretical tautology problem pointed out by van Knippenberg and Sitkin (2013) as the relationships between independent and dependent variables are semantically determined. One important finding however is that the residuals – or precision – of the predicted correlations is almost independent of the adjusted R^2 in each sample. The proportion of variance explained by semantics predicts only 3% of the variance in the residuals from linear regression from sample to sample. In other words, the non-semantic information is mostly noise, so that most of the signal is determined by the semantics – if there are relationships, these are most likely to be produced by semantics.

This is in line with hypothesis 2, which stated that "differences in survey response statistics between different ethnic and linguistic groups can be empirically explained by lack of understanding of the item texts." While this is not in itself a clear test of Hypothesis 2, this will be subjected to further testing below. However, we first want to test Hypothesis 3. This stated that "Samples of respondents who do not have English as their native language will display unique common variance that is neither explained by semantic algorithms nor by response patterns from unrelated cultures."

To identify the uniquely ethnic variance components in the data, we applied a stepwise hierarchical regression analysis, implying the following theoretical considerations: As argued initially, we assume that Chinese natives responding in Chinese will be most likely to display cultural differences from the native English speakers. We therefore enter the semantic similarity indices in the first block as the undisputedly semantic predictors of variance. As mentioned, the digital algorithms are still inferior to most adult human speakers in parsing semantic structures. In the second step, we therefore enter the values for native speakers of English. To the extent that these numbers express something in common with the native Chinese speakers, it should be something like the knowledge common to all humans with no special cultural significance. Further, we add Norwegians and Germans in their native languages in step 3, as there is no reason either to think that these groups share cultural characteristics with Chinese. In step 4, we add Norwegians and other Europeans in English. In step 5, we enter Pakistanis and Indian nationals in English, as we are now moving eastwards in cultural influence. In step 6, we enter non-Chinese East Asians in English. In step 7, we finally enter the Chinese responding in English. This allows us to inspect if the explained variance increases as we add samples with more Asian cultural elements. The result is displayed in **Table 4**.

Hypothesis 3 seems supported in that there is a unique component of variance comprising 5% that is shared only between the two Chinese samples responding in either Chinese or English.

However, the uniquely Chinese variance seems small. The bulk of variance seems predicted by the semantic algorithms alone (54%). Adding native English speakers and Europeans improve the prediction by 12%, reaching 66% with no probable influence from uniquely Chinese cultural heritage. There is an arguable Asian component in between – 3% from the Indian subcontinent or 4% from the non-Chinese East Asians.

The sample with the most deviant statistical pattern does however seem to be the Pakistanis responding in Urdu, not the Chinese as theoretically expected. We again tried the same stepwise regression to see if there is a uniquely Pakistani way of responding to the MLQ. As in the previous model, we entered the semantics and the native English speakers first. This time though, the Indian natives came toward the end, before the Pakistani sample in English was entered in the model.

As can be seen from **Table 5**, the uniquely Pakistani variance component (i.e., shared only between Pakistani respondents in Urdu and in English) is at most 3%. They do not share any unique variance at all with Indian natives.

To intensify the analysis of the seemingly aberrant statistics from Pakistanis in Urdu, we did a further breakdown of the dataset. 65 of the Urdu responses were collected in Pakistan and another 46 responses were collected among first generation immigrants to Norway. We repeated a stepwise regression model, entering only semantics and Pakistanis in English first, but this time tried to analyze how much unique variation the two different Urdu samples seemed to have. The results are displayed in **Table 6**, and it turns out that the two different Urdu samples have absolutely nothing uniquely in common.

TABLE 3 | Average correlations between leadership scales and the outcome measures, with their semantically predicted counterparts, by linguistic sub-sample.

Experimental group	Conditional reward	Individ. consideration	Idealized influence attr.	Idealized influence beh.	Inspiring motivation	Intellect. stimulation	Laissez-faire	Active mgmnt by except.	Passive mgmnt by except.	Outcome to outcome	All other relationships	Avg residuals
English native speakers	0.55	0.54	0.57	0.48	0.54	0.53	−0.45	0.26	−0.32	0.70	0.16	
Predicted in linear regr.	*0.48*	*0.52*	*0.48*	*0.45*	*0.48*	*0.46*	*−0.35*	*0.45*	*−0.34*	*0.56*	*0.17*	*0.07*
Predicted in GLM	*0.52*	*0.54*	*0.53*	*0.46*	*0.51*	*0.49*	*−0.39*	*0.33*	*−0.27*	*0.70*	*0.16*	*0.03*
Norwegians in Norwegian	0.47	0.54	0.52	0.51	0.52	0.50	−0.36	0.16	−0.19	0.60	0.18	
Predicted in linear regr.	*0.43*	*0.48*	*0.45*	*0.44*	*0.47*	*0.43*	*−0.25*	*0.42*	*−0.25*	*0.53*	*0.19*	*0.08*
Predicted in GLM	*0.47*	*0.52*	*0.48*	*0.49*	*0.50*	*0.47*	*−0.31*	*0.23*	*−0.16*	*0.60*	*0.18*	*0.03*
Norwegians in English	0.41	0.46	0.55	0.37	0.45	0.47	−0.37	−0.03	−0.26	0.63	0.13	
Predicted in linear regr.	*0.36*	*0.40*	*0.37*	*0.35*	*0.37*	*0.35*	*−0.23*	*0.34*	*−0.23*	*0.45*	*0.14*	*0.11*
Predicted in GLM	*0.37*	*0.44*	*0.44*	*0.35*	*0.39*	*0.40*	*−0.28*	*0.10*	*−0.16*	*0.63*	*0.13*	*0.06*
Germans in German	0.49	0.55	0.52	0.40	0.48	0.48	−0.41	0.15	−0.15	0.64	0.17	
Predicted in linear regr.	*0.39*	*0.45*	*0.42*	*0.41*	*0.45*	*0.40*	*−0.21*	*0.39*	*−0.20*	*0.49*	*0.18*	*0.10*
Predicted in GLM	*0.43*	*0.49*	*0.46*	*0.39*	*0.47*	*0.39*	*−0.28*	*0.24*	*−0.09*	*0.64*	*0.17*	*0.05*
Other Europeans in English	0.53	0.57	0.66	0.58	0.53	0.63	−0.57	0.08	−0.39	0.69	0.15	
Predicted in linear regr.	*0.50*	*0.54*	*0.50*	*0.46*	*0.47*	*0.48*	*−0.41*	*0.46*	*−0.41*	*0.59*	*0.16*	*0.11*
Predicted in GLM	*0.53*	*0.57*	*0.59*	*0.52*	*0.49*	*0.52*	*−0.49*	*0.22*	*−0.32*	*0.69*	*0.15*	*0.05*
Pakistanis in Urdu	0.18	0.20	0.28	0.18	0.22	0.25	0.19	0.12	0.08	0.08	0.38	
Predicted in linear regr.	*0.18*	*0.22*	*0.22*	*0.22*	*0.23*	*0.24*	*0.22*	*0.11*	*0.21*	*0.11*	*0.25*	*0.04*
Predicted in GLM	*0.20*	*0.26*	*0.17*	*0.21*	*0.23*	*0.16*	*0.11*	*0.16*	*0.08*	*0.38*	*0.18*	*0.09*
Pakistanis in English	0.14	0.35	0.34	0.44	0.30	0.46	0.30	−0.17	−0.12	−0.22	0.57	
Predicted in linear regr.	*0.15*	*0.26*	*0.32*	*0.30*	*0.30*	*0.32*	*0.28*	*−0.10*	*0.27*	*−0.10*	*0.36*	*0.11*
Predicted in GLM	*0.27*	*0.35*	*0.33*	*0.28*	*0.35*	*0.27*	*−0.09*	*0.04*	*−0.11*	*0.57*	*0.14*	*0.22*
Chinese in Chinese	0.37	0.33	0.34	0.38	0.42	0.45	−0.32	0.18	−0.10	0.53	0.18	
Predicted in linear regr.	*0.34*	*0.38*	*0.35*	*0.36*	*0.40*	*0.34*	*−0.13*	*0.35*	*−0.12*	*0.40*	*0.18*	*0.07*
Predicted in GLM	*0.35*	*0.32*	*0.32*	*0.38*	*0.39*	*0.38*	*−0.24*	*0.22*	*−0.08*	*0.53*	*0.18*	*0.03*
Chinese in English	0.42	0.33	0.41	0.39	0.41	0.43	−0.26	0.19	−0.22	0.56	0.16	
Predicted in linear regr.	*0.37*	*0.37*	*0.34*	*0.34*	*0.36*	*0.34*	*−0.16*	*0.34*	*−0.16*	*0.40*	*0.16*	*0.07*
Predicted in GLM	*0.39*	*0.31*	*0.37*	*0.36*	*0.37*	*0.38*	*−0.18*	*0.23*	*−0.13*	*0.56*	*0.16*	*0.04*
Indian natives in English	0.51	0.45	0.63	0.44	0.58	0.52	−0.47	0.25	−0.29	0.61	0.17	
Predicted in linear regr.	*0.49*	*0.51*	*0.47*	*0.44*	*0.45*	*0.45*	*−0.33*	*0.45*	*−0.33*	*0.54*	*0.17*	*0.08*
Predicted in GLM	*0.51*	*0.48*	*0.57*	*0.42*	*0.53*	*0.47*	*−0.42*	*0.32*	*−0.25*	*0.61*	*0.17*	*0.03*

(Continued)

TABLE 3 | Continued

Experimental group	Conditional reward	Individ. consideration	Idealized influence attr.	Idealized influence beh.	Inspiring motivation	Intellect. stimulation	Laissez-faire	Active mgmnt by except.	Passive mgmnt by except.	Outcome to outcome	All other relationships	Avg residuals
East Asians non-Chinese English	0.46	0.42	0.46	0.52	0.53	0.46	−0.11	0.13	0.06	0.58	0.26	
Predicted in linear regr.	*0.39*	*0.44*	*0.42*	*0.43*	*0.46*	*0.40*	*0.00*	*0.40*	*0.01*	*0.47*	*0.26*	*0.08*
Predicted in GLM	*0.40*	*0.43*	*0.45*	*0.48*	*0.50*	*0.42*	*−0.09*	*0.22*	*0.10*	*0.58*	*0.26*	*0.03*
Whole dataset correlations $N = 2513$	0.18	0.44	0.45	0.48	0.44	0.49	0.47	−0.31	0.13	−0.18	0.59	
Predicted in linear regr.	*0.18*	*0.39*	*0.43*	*0.40*	*0.40*	*0.43*	*0.39*	*−0.20*	*0.38*	*−0.19*	*0.48*	*0.07*
Predicted in GLM	*0.18*	*0.42*	*0.44*	*0.43*	*0.42*	*0.45*	*0.42*	*−0.26*	*0.21*	*−0.14*	*0.59*	*0.03*

As argued in our initial discussion, we suspected that something was wrong with the Urdu translation of the MLQ, or with the samples, and we therefore contacted a certified Urdu translator who judged the materials. He could quickly give us a likely explanation for the chaotic statistics. Many Pakistani citizens will actually not have Urdu, but Punjabi as their native language. However, while Urdu is also a written language, Punjabi is only an oral language, a fact corroborated by a linguistic report on Pakistanis in Norway (Thiesen, 2003). Many Pakistanis will therefore claim that their native language is Urdu, even if this is strictly not correct. The most likely reason for the noisy statistical patterns is therefore simply a lack of understanding – the respondents have inadequate reading skills in Urdu. We take this as support for hypothesis 2, claiming that lack of linguistic proficiency is the most likely cause of reduced semantic predictability where this is elsewhere found to be substantial. A further corroboration of this interpretation can be found by comparing the Norwegians responding in Norwegian to the Norwegians responding in English. Since the English survey version among English native speakers is the most semantically predictable condition, the lack of semantic predictability of Norwegians is probably due to the difference in their proficiency in English and their native language. Lack of proficiency in English is the best explanation for the drop in semantic predictability.

As a final check, we subjected all the 990 item pair correlations for each linguistic sub-sample with the semantic values to a PCA with varimax rotation. This is a procedure used earlier to separate and map languages and genes according to anthropological developments, and tends to yield clusters of related languages (Cavalli-Sforza, 2001). The PCA displayed two factors, displayed as a 2-factor plot in **Figure 1**. It can be seen that one factor is essentially made up of the sample responding in Urdu. The rest of the sample clusters unsystematically around the semantic values created by the algorithms. Thus, there are no signs that the responses in Urdu are culturally determined, but most likely a result of inadequate language skills. Also, the two-dimensional plot supports H1 in that the overwhelming determinant of variation in the data is semantic.

DISCUSSION

The purpose of this study was to explore the extent to which semantic algorithms can predict correlation matrices across different languages and national samples in a semantically determined leadership survey. It was theorized that the propositional structures inherent in semantic information are largely translatable across languages.

This study administered a globally prevalent leadership survey with established semantic properties to a broad cross-cultural sample spanning the Anglo-Saxon cultural domain (native English speakers), northern Europe (Norwegians and Germans), the Indian subcontinent (Pakistani and Indian natives) as well as East Asia (China, Korea, Indonesia, Malaysia and Japan).

TABLE 4 | Predicting Chinese outcome patterns in hierarchical regression by semantics and other subgroups.

Cultural influence	Models	Adjusted R^2	Adjusted R^2 increase	df	Mean square	F
Algorithm block	(1) Semantic algorithms alone	0.54		4	12.03	287.57
European language block	(2) Adding native English speakers	0.63	0.09	5	11.28	337.18
	(3) Adding Norwegians and Germans in their native languages	0.64	0.01	7	8.17	250.11
	(4) Adding Norwegians and other Europeans in English	0.66	0.02	9	6.59	215.08
Indian subcontinent	(5) Adding Indian and Pakistani natives in English	0.69	0.03	11	5.60	197.79
East Asian	(6) Adding non-Chinese East Asians in English	0.73	0.04	12	5.44	221.70
Uniquely Chinese	(7) Adding Chinese in English	0.77	0.05	13	5.34	261.09

P-values for all models and increases in $R^2 < 0.001$.

For all sub-samples, the semantic algorithms predicted significant proportions of the variation in correlations between items, ranging from 11 to 84%. The semantic algorithms were computed using the English version of a survey originating from the United States. It is therefore natural that the best predicted sample was the native speakers of English (mostly United States and United Kingdom citizens).

The next best prediction occurred also mostly for samples responding in English. This was true for non-English speakers from Europe, Indian nationals and even Chinese respondents in English. The differences in statistical patterns are therefore largely attributable to linguistic precision and understanding. One interesting example is provided by the two Norwegian samples. Norwegian is a Germanic language close to English (Renfrew, 1987; Cavalli-Sforza, 2001), and Norwegians are usually quite competent speakers of English (Warner-Søderholm, 2013). There is no wonder therefore, that both samples seem fairly semantically determined. As the Norwegian sample responding in English is slightly less semantically predictable than the one responding in their native language may therefore be due to a lack of linguistic precision. The Norwegian language version of the MLQ may be quite adequate, and better than the private translation that a Norwegian respondent needs to do while responding to a version in English.

For samples with a more remote relationship to English, there may be other explanations. Chinese and Pakistani nationals respond much more semantically driven when responding to an English version of the scales than to versions in their own native languages.

To the extent that survey data are sensitive to cultural differences, such effects should arguably be most likely to occur in the non-English speaking samples responding in their native languages (Boroditsky, 2011; Harzing et al., 2011; Zander et al., 2011). Hence, it is natural that the semantic algorithms show lesser predictive values for these than to other samples. However, it is hard to say what these differences in response patterns may imply (cfr. Russell, 1922; Behr et al., 2016). A comparison of the predicted correlations with the observed ones showed that these were fairly close even in the case where semantics predicted only weakly. This is a finding akin to earlier findings in studies of variance components in responses and semantic predictability, in that the semantic patterns are the main driver of the observed correlations (Arnulf et al., 2018b). If other variance components exerted notable influence, the English language semantic values should be systematically off target the more culturally disparate the sub-sample was. This did not seem to be the case. Among the respondents in the sample, there were obviously groups very different from the native English speakers. Still, the response

TABLE 5 | Predicting Pakistani outcome patterns in hierarchical regression.

Cultural influence	Models	Adjusted R^2	Adjusted R^2 increase	df	Mean square	*F*	
Algorithm block	(1) Semantic algorithms alone	0.11		4	0.66	31,92	
European language block	(2) Adding native English speakers	0.20	0.09	5	0.94	5.25	
	(3) Adding Norwegians and Germans in their native languages	0.20	0.00	7	0.68	36,60	ns
	(4) Adding Norwegians and other Europeans in English	0.25	0.05	9	0.64	35,85	
East Asian	(5) Adding Chinese in Chinese	0.25	0.00	10	0.57	32,25	ns
	(6) Adding Chinese and non-Chinese East Asians in English	0.26	0.01	12	0.49	27,85	
Indian subcontinent	(7) Adding Indian Natives in English	0.26	0.00	13	0.46	25,83	ns
Uniquely Pakistani	(8) Adding Pakistanis in English	0.29	0.03	14	0.48	28,12	

TABLE 6 | The Urdu samples from Pakistan and Norway in hierarchical regression.

Cultural influence	Models	Adjusted R^2	Adjusted R^2 increase	df	Mean square	*F*	
Algorithm block	(1) Semantic algorithms alone	0.03		4	0.36	8,0.0	
European language block	(2) Adding native English speakers	0.05	0.02	5	0.50	11.86	
Pakistanis in English	(3) Adding Pakistanis in English	0.06	0.01	6	0.50	11.99	
Uniquely Urdu	(4) Adding Pakistanis from Norway in Urdu	0.06	0.00	7	0.44	10.41	ns

patterns were notably influenced by semantics as predicted by the algorithms.

The strongest deviations from the semantic patterns were found in the Pakistani sample responding in Urdu. The two Urdu samples, the one in Pakistan and the one in Norway, had no shared variation, and did not share unique variation with either other Pakistanis in English or the sub-sample from the Indian subcontinent that would be their most likely cultural relative. Everything considered, the statistics in the Urdu samples were most likely influenced by problems with the translation of the survey and even more by inadequate reading capabilities in the respondents. This is also in line with other research that has replicated the variable structure of transformational leadership in Pakistan (Khan et al., 2014).

This study made the theoretical claim that Chinese responding in Chinese should appear as culturally most distant to the native English speakers. If we disregard the obvious language problem in the Urdu group, the Chinese responding in Chinese did display the lowest semantic predictability in the study, as expected. However, when we controlled for all non-Chinese

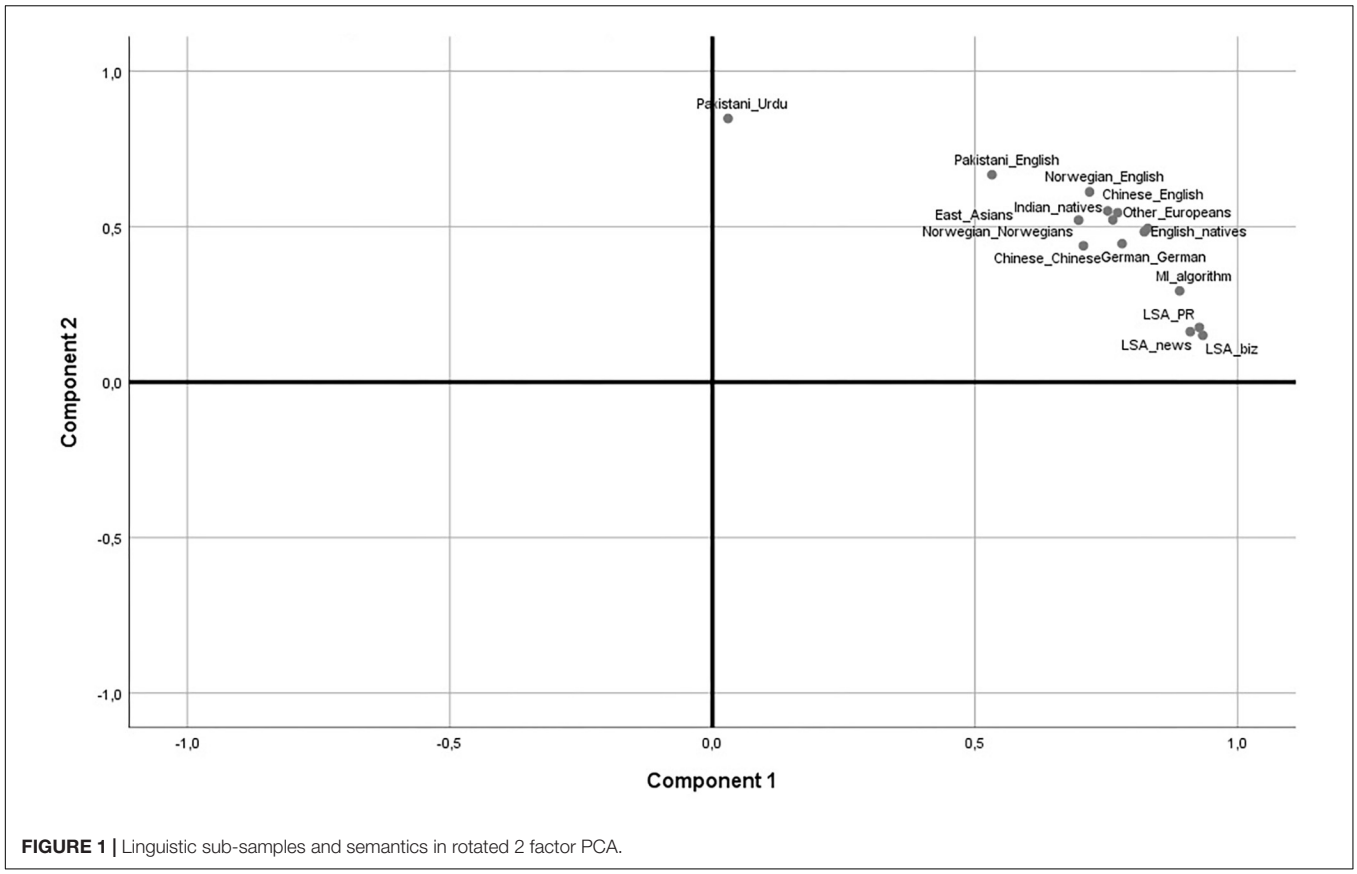


FIGURE 1 | Linguistic sub-samples and semantics in rotated 2 factor PCA.

speakers, there was not much unique variation left among the Chinese respondents. The two Chinese samples responding in English and Chinese shared only 5% unique variation, less than a tenth of the variation they shared with the numbers from the digital algorithms. The unique variations between the ethnic samples in the native/English conditions were always around one to five percent, which may well be within random range. This shared variation was of the same magnitude as the differences within the non-Asian samples and within Asia. There are no compelling reasons to attribute these differences to cultural similarities between Chinese and Indians, or between Chinese and Japanese for that matter (Wang and Satow, 1994; Liu et al., 2004; Aoki, 2008).

A recent study on significant differences between score levels of groups has indicated that even with notable *p*-values and effect sizes, similarities in group distributions may practically outweigh the noted difference substantially (Hanel et al., 2019). The study proposes a measure called absolute effect (AE), defined as the median difference expressed as the percentage of the largest possible scale difference. Exploring the Semantic Theory of Survey Response we are usually not concerned with the score levels *per se*. Instead, we are investigating how the mutual patterns among survey responses reflect semantically given structures. If we apply the rationale behind the AE on the semantic structure in our study, a 5% shared unique variance among Chinese respondents equals an average "freedom" in responses in the MLQ of 5% of at most 0.2 scale points on a 5-scale Likert scale option. Or stated differently, the median Chinese respondent may be expected to depart 0.2 Likert scale score points from an English native speaker. The practical impact of this is hard to grasp in terms of measurement theory (McGrane and Maul Gevirtz, 2019).

From the earlier studies in this field, we know that the semantic structure usually emerges quite quickly with even a few respondents when it is as salient as in the present instrument (Arnulf et al., 2014a, 2018b). Sample sizes do not seem to be very crucial above a certain level. In the present case, the semantics predicted about equally well in the huge sample of Norwegians in Norwegian as in the much smaller samples such as Germans in German and English Natives. As expected, the Chinese samples seem to require a few more respondents for the matrix to approach the semantically given values. If some of our samples are below the optimal threshold for semantic predictability, increasing sample sizes would most likely increase the fit between semantic and respondent matrices.

Previous research has also indicated that groups of respondents display variance components from many sources, including personality and management level (Arnulf et al., 2018b). This is in accordance with what is expected from other studies on respondent characteristics in cross-cultural research (Harzing, 2006; He et al., 2014). This line of research asserts that differences between culturally divergent groups cannot be

attributed to culture unless their respondent characteristics are controlled and accounted for. Our perspective is the opposite – we are simply aiming to show how much semantic patterns will unite proposedly different groups. Since our focus is on the extent of semantic influences, and since teasing apart variance components from the semantic structures requires more extensive laboratory work, this study has refrained from decomposing the origins of semantic structures further.

Taken together, our findings raise questions about the value of semantically driven surveys as a tool in cross-cultural leadership research methods. We believe that our data warrant the following three conclusions:

Semantically Determined Surveys May Be Insensitive to Cultural Differences

The replication of data structures from semantically determined surveys may not tell us much about cultural differences, except for the fact that propositional structures in the survey have been correctly translated. This is a failure to distinguish between logical and empirical research questions (Russell, 1918/2007, 1922; Semin, 1989; Lovasz and Slaney, 2013; Smedslund, 2015; Arnulf et al., 2018b). The answers to logical research questions are given *a priori*, which is the reason why the response statistics are predictable by using computer algorithms that know nothing about respondents or cultures. This kind of research risks asserting that people and organizations are the same everywhere, disregarding the participants' experiences that leadership phenomena are actually quite different across contexts (Henrich et al., 2010; Mendenhall, 2013). It is also likely to inflate statistics in ways that have frequently been demonstrated as effects of common method variance (Podsakoff et al., 2012; Schaller et al., 2015).

Equivocality of Non-replication

Conversely, the main reason for observed differences in cases like the one we study here may simply be linguistic problems, either in the translation or in the respondents' decoding of the item texts (Behr et al., 2016). The differences between samples in this study show that while the big bulk of relationships are semantically driven, there may be detectable differences that can masquerade as cultural differences because they are linked to different linguistic and ethnical groups. However, our findings also show that these differences may easily be explained by lack of language skills, local interpretations or faulty interpretations of the survey instrument. Even small differences in interpretations seem to influence the response statistics.

Cross-Cultural OB Research Needs Better Philosophical Groundwork

The use of surveys in cross-cultural research on OB has for years avoided dealing with the difficult topic of what the "measurements" actually measure (Smedslund, 1988; Drasgow et al., 2015; Maul, 2017; Slaney, 2017). The original assumption of Likert (1932) was that the scales measure attitude strength, and that the ensuing statistical patterns were indicative of behavioral dispositions or inclinations. This assumption was originally doubted by his contemporaries in psychometrics, but Likert's views prevailed as increasingly sophisticated statistical tools offered hopes of mathematical refinement (van Schuur and Kiers, 1994; Andrich, 1996). In recent years, though, the assumptions underlying measurements have come under renewed scrutiny. Some of the core psychometric criteria for construct validation are not capable of falsifying erroneous hypotheses, and the "measurements" may be measuring quite different entities from what they purport (Slaney and Racine, 2013; Mari et al., 2017; Maul, 2017; Arnulf et al., 2018b; Kjell et al., 2019).

The lack of awareness about these problems is all the more unfortunate in cross-cultural leadership, due to the risk of ethnocentrism inherent in the core problems of this field (Ng et al., 2009; Zhang et al., 2014; Ma and Tsui, 2015; Nagai et al., 2015). There is growing documentation about the fact that scholars as well as research subjects from "WEIRD" (White, Educated, Industrialized, Rich, and Democratic) countries are overwhelmingly represented with subsequent risks of theoretical and empirical biases (Henrich et al., 2010; Hibbing et al., 2014). Cross-cultural leadership is of great practical relevance in business and politics, and the costs of failures in this field are probably large (Gutierrez et al., 2012; Porter and Rivkin, 2012; Osland et al., 2013; Arvey et al., 2015). Anthropologists have for decades warned against the use of "thin data" in research on cross-cultural topics (Geertz, 1973).

When constructs like leadership are found to be semantically predictable to the extent found in this case, the most likely theoretical explanation is that it is precisely socially constructed (Berger and Luckman, 1966; Grint, 2005; Fairhurst and Grant, 2010). In this case, items may not so much be empirical "measures" as they are defining characteristics of the social construction (Smedslund, 1988; Elster, 2011, 2018; Lovasz and Slaney, 2013; Maul, 2017). The inter-item correlation matrix will then most likely reflect these mutual patterns in most languages whether the social construct is adopted in that culture or not.

The specific conclusion from this study is that cross-cultural studies in leadership need a more sophisticated view on the relationship between language and action in theory as well as practice. Studies that pick up semantic patterns are more likely to be language research than research on actions, a difference dealt with at length in action theory and control theory (Frese and Zapf, 1994; Weseman, 2007; Prinz et al., 2009; Parks-Stamm et al., 2010; Schaller et al., 2015; Gantman et al., 2017). When response patterns from semantically driven surveys are replicable across contexts, it may only mean that the same sentences can be said, with approximately the same understanding, across these contexts. This is unsurprising in itself – it equals the mere methodological requirement to have surveys translated and re-translated to ensure their identical meaning across languages (Herdman et al., 1997). In today's global economies, most sentences that describe working environments may be translated from one language to another.

That is not the same as saying that the same things matter, that acts are carried out the same way, and with the same effects on people in the surroundings. The epistemological error that

seems to be frequently committed in organizational behavior is to confuse behaviors with their intentions and effects on an abstract level. This has been theoretically proven by van Knippenberg and Sitkin (2013) in the case of transformational leadership, where definitions and operationalizations conflate independent and dependent variables.

Recent developments in indigenous Chinese research on leadership shows the likelihood that there exist distinct types of leadership behaviors that also have distinct effects on Chinese employees. This differs from the effects on, e.g., Western employees in the same companies (Chen and Kao, 2009; Cheng et al., 2014; Chen et al., 2015; Qin et al., 2015; Zhang et al., 2015). We obviously need more efforts to address the perceived differences that practitioners and scholars alike experience in the field, and generate instruments that capture these differences instead of neglecting them. That requires a less ethnocentric and more advanced philosophical foundation for understanding the role of language in research and cross-cultural leadership.

LIMITATIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The present study is a cross-sectional analysis of the responses to one single type of leadership survey. We believe that this is warranted, as we do not look at the temporal effects of the responses, but simply at the degree to which they are semantically determined. This means that the independent variable – the output from semantic algorithm – is not conflated with the dependent variables, i.e., the human responses. Also, we believe that the MLQ is an important exemplary type of leadership survey as it has been analyzed for its semantic structure in earlier publications and is a common instrument in cross-cultural leadership research.

The present study uses a series of mixed samples of various sizes and from various industries, locations and cultures. One clear limitation of our design is that the sub-samples are of unequal size and they are also not matched in terms of demographics and educational characteristics. We have no stringent control over the "cultural" diversity in the samples except for the locations and the languages of the respondents. We do think that our design goes a long way to randomize factors like industries, professions and other non-intended sample characteristics. Still, there may be better methods to control and document the cultural conditions that are central in determining differences in leadership.

One particular limitation of the present study is that we have only used English language items to inform the algorithms. As expected, the ability of the algorithms to predict response patterns were better for English and linguistically related groups than for groups with cultures and languages more distant to English. Our design can for good reason be suspected of adopting a culturally skewed perspective in the algorithms themselves. As explained, we believe this is warranted as a first step here due to the WEIRD heritage of the leadership constructs and measurement instruments themselves. The semantic perspective raises a question about how indigenous, non-WEIRD leadership issues should be conceptualized both as theoretical constructs and as measurements. Further developments in this field are necessary to create a viable research agenda here.

Finally, this study did not look specifically at cultural differences in score levels between cultures. We do think that more valid information about cross-cultural leadership research can be found in that direction. This study has concentrated on studying the relationships between item pairs and subscales, as these are frequently used as important inputs for further statistical modeling.

For future research, we highly recommend more controlled studies where the semantic influences on survey statistics are more clearly identified as sources of variation. We know that attempts at using multi-trait multi-method approaches are under way (Martinsen et al., 2017). It is imperative that the semantic components are identified and properly understood, for example as sources of common method variance (Bagozzi, 2011) or as a general response style (He et al., 2014). To truly understand the unique impact of semantic relationships in cross-cultural research, we nee more knowledge about high-quality instruments with balanced items, so that the effect of item types on the semantic structure would be easier to discern.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Norsk Samfunnsvitenskapelig Datatjeneste. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JA designed the study, collected the survey data materials, and co-wrote the manuscript. KL provided the semantic algorithms, helped develop the theory, and co-wrote the manuscript.

ACKNOWLEDGMENTS

We want to express our gratitude to Dr. Øyvind Martinsen, Lily Chen, Aisha Qureshi, and Lisa Stengel for helping with the data collection.

REFERENCES

AERA, APA, and NCME, (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling thurstone and likert methodologies. *Br. J. Math. Statist. Psychol.* 49, 347–365. doi: 10.1111/j.2044-8317.1996.tb01093.x

Aoki, K. (2008). Transferring Japanese kaizen activities to overseas plants in China. *Int. J. Operat. Product. Manag.* 28, 518–539. doi: 10.1108/01443570810875340

Arnulf , J. K. (2020). "Wittgenstein's revenge: how semantic algorithms can help survey research escape smedslund's labyrinth," in *Respect for Thought; Jan Smedslund's Legacy for Psychology*, eds T. G. Lindstad, E. Stänicke, and J. Valsiner, (Berlin: Springer).

Arnulf, J. K., and Larsen, K. R. (2015). Overlapping semantics of leadership and heroism: expectations of omnipotence, identification with ideal leaders and disappointment in real managers. *Scand. Psychol.* 2:e3. doi: 10.15714/scandpsychol.2.e3

Arnulf, J. K., Larsen, K. R., and Dysvik, A. (2018a). Measuring semantic components in training and motivation: a methodological introduction to the semantic theory of survey response. *Hum. Resour. Dev. Q.* 30, 17–38. doi: 10.1002/hrdq.21324

Arnulf, J. K., Larsen, K. R., Martinsen, ØL., and Egeland, T. (2018b). The failing measurement of attitudes: how semantic determinants of individual survey responses come to replace measures of attitude strength. *Behav. Res. Methods* 50, 2345–2365. doi: 10.3758/s13428-017-0999-y

Arnulf, J. K., Larsen, K. R., and Martinsen, ØL. (2018c). Respondent robotics: simulating responses to likert-scale survey items. *Sage Open* 8, 1–18. doi: 10.1177/2158244018764803

Arnulf, J. K., Larsen, K. R., and Martinsen, ØL. (2018d). Semantic algorithms can detect how media language shapes survey responses in organizational behaviour. *PLoS One* 13:e0207643. doi: 10.1371/journal.pone.0207643

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Bong, C. H. (2014a). Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. *PLoS One* 9:e106361. doi: 10.1371/journal.pone.0106361

Arnulf, J. K., Niu, Z., and Lu, H. (2014b). Management, culture and language: native and second languages in management team communication. *Paper Presented at the Academy of Management Meeting*, Philadelphia.

Arnulf, J. K., Larsen, K. R., Martinsen, ØL., and Bong, C. H. (2015). A new approach to psychological measures in leadership research. *Scand. Psychol.* 2:e4. doi: 10.15714/scandpsychol.2.e4

Arvey, R., Dhanaraj, C., Javidan, M., and Zhang, Z.-X. (2015). Are there unique leadership models in Asia? Exploring uncharted territory. *Leadersh. Q.* 26, 1–6. doi: 10.1016/j.leaqua.2015.01.003

Avolio, B. J., Bass, B. M., and Jung, D. I. (1995). *Multifactor Leadership Questionnaire Technical Report*. Redwood City, CA: Mind Garden.

Baeza-Yates, R. (2018). Bias on the web. *Commun. ACM* 61, 54–61. doi: 10.1145/3209581

Bagozzi, R. P. (2011). Measurement and meaning in information systems and organizational research: methodological and philosophical foundations. *Mis. Q.* 35, 261–292. doi: 10.2307/23044044

Bass, B. M. (1997). Does the transactional-transformational leadership paradigm transcend organizational boundaries? *Am. Psychol.* 52, 130–139.

Bass, B. M. (1998). *Transformational Leadership: Industry, Military, and Educational Impact*. Mahwah, NJ: Erlbaum.

Behr, D., Braun, M., and Dorer, B. (2016). *Measurement Instruments in International Surveys. GESIS Survey Guidelines*. Mannheim: GESIS – Leibniz Institute for the Social Science.

Berger, P., and Luckman, T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Garden City, NY: Doubleday.

Boroditsky, L. (2001). Does language shape thought: mandarin and English speakers' conceptions of time. *Cogn. Psychol.* 43, 1–22.

Boroditsky, L. (2011). *How Language Shapes Thought:The languages We Speak Affect Our Perceptions of the World*. Berlin: Springer, 43–45.

Boroditsky, L., and Gaby, A. (2010). Remembrances of times East: absolute spatial representations of time in an Australian aboriginal community. *Psychol. Sci.* 21, 1635–1639. doi: 10.1177/0956797610386621

Borsboom, D. (2008). Latent variable theory. *Measurement* 6, 25–53.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *J. Cross Cult. Psychol.* 1, 185–216. doi: 10.1177/135910457000100301

Cavalli-Sforza, L. L. (2001). *Genes, Peoples, and Languages*. London: Penguin Books.

Chen, H. Y., and Kao, H. S. R. (2009). Chinese paternalistic leadership and non-Chinese subordinates' psychological health. *Int. J. Hum. Resour. Manag.* 20, 2533–2546. doi: 10.1080/09585190903363839

Chen, L., Yang, B. Y., and Jing, R. T. (2015). Paternalistic leadership, team conflict, and TMT decision effectiveness: interactions in the chinese context. *Manag. Organ. Rev.* 11, 739–762. doi: 10.1017/mor.2015.34

Cheng, B. S., Boer, D., Chou, L. F., Huang, M. P., Yoneyama, S., Shim, D., et al. (2014). Paternalistic leadership in four east asian societies: generalizability and cultural differences of the triad model. *J. Cross Cult. Psychol.* 45, 82–90. doi: 10.1177/0022022113490070

Costa, A., Vives, M. L., and Corey, J. D. (2017). On language processing shaping decision making. *Curr. Direct. Psychol. Sci.* 26, 146–151. doi: 10.1177/0963721416680263

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inform. Sci.* 41, 391–407.

Dennis, S., Landauer, T. K., Kintsch, W., and Quesada, J. (2013). *Introduction to Latent Semantic Analysis*. Boulder: University of Colorado.

Drasgow, F., Chernyshenko, O. S., and Stark, S. (2015). 75 years after likert: thurstone was RIGHT! *Indust. Organ. Psychol.* 3, 465–476. doi: 10.1111/j.1754-9434.2010.01273.x

Dumdum, U. R., Lowe, K. B., and Avolio, B. C. (2002). "A meta analysis of the transformational and transactional leadership correlates of effectiveness and satisfaction: an update and extension," in *Transformational and Charismatic Leadership: The Road Ahead*, eds B. J. Avolio, and F. J. Yammarino, (Amsterdam: JAI Press), 35–66.

Elster, J. (2011). Hard and soft obscurantism in the humanities and social sciences. *Diogenes* 58:159. doi: 10.1177/0392192112444984

Elster, J. (2018). "How my theory explains everything: and can make you happier, healthier, and wealthier," in *Hanging on to the Edges: Essays on Science, Society and the Academic Life*, ed. D. Nettle, (Cambridge: Open Book Publishers).

Enfield, N. J. (2007). "Tolerable friends," in *Proceedings of the 33rd Annual Meeting of the Berkeley Linguistics Society*, (Berkeley, CA: Berkeley Linguistics Society).

Fairhurst, G. T., and Grant, D. (2010). The social construction of leadership: a sailing guide. *Manag. Commun. Q.* 24, 171–210. doi: 10.1177/0893318909359697

Fausey, C. M., Long, B. L., Inamori, A., and Boroditsky, L. (2010). Constructing agency: the role of language. *Front. Psychol.* 1:162. doi: 10.3389/fpsyg.2010.00162

Feldman, J. M., and Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *J. Appl. Psychol.* 73, 421–435. doi: 10.1037//0021-9010.73.3.421

Frese, M., and Zapf, D. (1994). "Action as the core of work psychology: a German approach," in *Handbook of Industrial and Organizational Psychology*, Vol. 4, eds H. C. Triandis, M. D. Dunnette, and L. M. Hough, (Palo Alto, CA: Consulting Psychologists Press), 271–340.

Gantman, A. P., Adriaanse, M. A., Gollwitzer, P. M., and Oettingen, G. (2017). Why did I do that? explaining actions activated outside of awareness. *Psychon. Bull. Rev.* 24, 1563–1572. doi: 10.3758/s13423-017-1260-5

Geertz, C. (1973). *The Interpretation of Cultures*. New York, NY: Basic Books.

Gefen, D., Endicott, J. E., Miller, J., Fresneda, J. E., and Larsen, K. R. (2017). A guide to text analysis with latent semantic analysis in r with annotated code: studying online reviews and the stack exchange community. *Commun. Assoc. Inform. Syst.* 41, 450–496.

Gefen, D., and Larsen, K. (2017). Controlling for lexical closeness in survey research: a demonstration on the technology acceptance model. *J. Assoc. Inform. Syst.* 18, 727–757. doi: 10.17705/1jais.00469

Gentner, D. (2016). Language as cognitive tool kit: how language supports relational thought. *Am. Psychol.* 71, 650–657. doi: 10.1037/amp0000082

Gesteland, R. R. (2012). *Cross-Cultural Business Behavior: A Guide for Global Management*. Køge: Copenhagen Business School Press.

Grint, K. (2005). Problems, problems, problems: the social construction of "leadership.". *Hum. Relat.* 58, 1467–1494.

Gumperz, J., and Levinson, S. C. (1996). "Introduction: linguistic relativity re-examined," in *Rethinking Linguistic Relativity*, eds J. Gumperz, and S. C. Levinson, (Cambridge: Cambridge University Press), 1–20.

Gutierrez, B., Spencer, S. M., and Zhu, G. R. (2012). Thinking globally, leading locally: chinese, Indian, and Western leadership. *Cross Cult. Manag. Int. J.* 19, 67–89. doi: 10.1108/13527601211195637

Hanel, P. H. P., Maio, G. R., and Manstead, A. S. R. (2019). A new way to look at the data: similarities between groups of people are large and important. *J. Pers. Soc. Psychol.* 116, 541–562. doi: 10.1037/pspi0000154.supp

Hanks, W. F. (1996). "Language form and communicative practices," in *Rethinking Linguistic Relativity*, eds J. Gumperz, and S. C. Levinson, (Cambridge: Cambridge University Press), 232–270.

Hansen, M. T., Ibarra, H., and Peyer, U. (2013). The best-performing CEOs in the world. *Har. Bus. Rev.* 91, 81–95.

Harzing, A. W. (2006). Response styles in cross-national survey research: a 26-country study. *Int. J. of Cross Cul. Manag.* 6, 243–266. doi: 10.1177/1470595806066332

Harzing, A. W., Koster, K., and Magner, U. (2011). Babel in business: the language barrier and its solutions in the HQ-subsidiary relationship. *J. World Bus.* 46, 279–287. doi: 10.1016/j.jwb.2010.07.005

He, J., Van de Vijver, F. J., Espinosa, A. D., and Mui, P. H. (2014). Toward a unification of acquiescent, extreme, and midpoint response styles: a multilevel study. *Int. J. Cross Cult. Manag.* 14, 306–322. doi: 10.1177/1470595814541424

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behav. Brain Sci.* 33, 61–135. doi: 10.1017/S0140525X0999152X

Herdman, M., Fox-Rushby, J., and Badia, X. (1997). 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. *Qual. Life Res.* 6, 237–247. doi: 10.1023/a:1026410721664

Hibbing, J. R., Smith, K. B., and Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *Behav. Brain Sci.* 37, 297–307. doi: 10.1017/S0140525X13001192

Hofstede, G., Hofstede, G. J., and Minkov, M. (2010). *Cultures and Organizations: Software Of The Mind: Intercultural Cooperation And Its Importance For Survival*. New York, NY: McGraw-Hill.

House, R., Hanges, P., Javidan, M., Dorfman, P., and Gupta, V. (eds) (2004). *Culture, Leadership, and Organizations: The Globe Study of 62 Societies*. Thousand Oaks: Sage.

Jiang, J. J., and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *Paper presented at the International Conference on Computational Linguistics*, Taiwan.

Jöreskog, K. G. (1993). "Testing structural equation models," in *Testing Structural Equation Models*, eds K. A. Bollen, and J. S. Long, (Newbury Park: Sage), 294–316.

Khan, M. I., Awan, U., Yasir, M., Bin Mohamad, N. A., Shah, S. H. A., Qureshi, M. I., et al. (2014). Transformational leadership, emotional intelligence and organizational commitment: pakistan's services sector. *Argum. Oecon.* 33, 67–92.

Kirkman, B. L., Lowe, K. B., and Gibson, C. B. (2006). A quarter century of culture's consequences: a review of empirical research incorporating Hofstede's cultural values framework. *J. Int. Bus. Stud.* 37, 285–320. doi: 10.1057/palgrave.jibs.8400202

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikstrom, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal About the Mind*. Chicago: The University of Chicago Press.

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: what is it? how was it contracted? Is there a cure? *New Ideas Psychol.* 31, 65–71. doi: 10.1016/j.newideapsych.2011.02.009

Landauer, T. K. (2007). "LSA as a theory of meaning," in *Handbook of Latent Semantic Analysis*, eds T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsh, (Mahwah, NJ: Larence Erlbaum Associates, Publishers), 3–34.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Process.* 25, 259–284. doi: 10.1080/01638539809545028

Larsen, K. R., and Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Q.* 40:529. doi: 10.25300/Misq/2016/40.3.01

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* 140, 1–55.

Lin, D. (1998). An information-theoretic definition of similarity. *Paper Presented at the 15th International Conference on Machine Learning*, Madison, WI.

Liu, C. H., Tjosvold, D., and Wong, M. (2004). Effective Japanese leadership in China: co-operative goals and applying abilities for mutual benefit. *Int. J. Hum. Resour. Manag.* 15, 730–749. doi: 10.1080/0958519042000192924

Lovasz, N., and Slaney, K. L. (2013). What makes a hypothetical construct "hypothetical"? Tracing the origins and uses of the 'hypothetical construct' concept in psychological science. *New Ideas Psychol.* 31, 22–31. doi: 10.1016/j.newideapsych.2011.02.005

Lucy, J. (1996). "The scope of linguistic relativity: an analysis and a review of empirical research," in *Rethinking Linguistic Relativity*, eds J. Gumperz, and S. C. Levinson, (Cambridge: Cambridge University Press), 37–69.

Ma, L., and Tsui, A. S. (2015). Traditional Chinese philosophies and contemporary leadership. *Leadersh. Q.* 26, 13–24. doi: 10.1016/j.leaqua.2014.11.008

MacKenzie, S. B., Podsakoff, P. M., and Podsakoff, N. P. (2011). Construct measurement and validation procedures in mis and behavioral research: integrating new and existing techniques. *MIS Q.* 35, 293–334.

Mansour, M. J. J., House, R. J., Dorfman, P. W. P. D., Hanges, P. J. P. H., and Luque, M. S. S. (2006). Conceptualizing and measuring cultures and their consequences: a comparative review of GLOBE's and Hofstede's approaches. *J. Int. Bus. Stud.* 37, 897–914.

March, J. G., and Sutton, R. I. (1997). Organizational performance as a dependent variable. *Organ. Sci.* 8, 698–706. doi: 10.1287/orsc.8.6.698

Mari, L., Maul, A., Irribarra, D. T., and Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement* 100, 115–121. doi: 10.1016/j.measurement.2016.12.050

Martinsen, ØL., Arnulf, J. K., Larsen, K. R., Ohlsson, U. H., and Satorra, A. (2017). Semantic influence on the measurement of leadership: a multitrait-multisource perspective. *Paper Presented at the Academy of Management Meeting*, Atlanta.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measur. Interdiscipl. Res. Perspect.* 15, 51–69. doi: 10.1080/15366367.2017.1348108

McEachrane, M. (2009). Emotion, meaning, and appraisal theory. *Theory Psychol.* 19, 33–53. doi: 10.1177/0959354308101418

McGrane, J. A., and Maul Gevirtz, A. (2019). The human sciences, models and metrological mythology. *Measurement* 152:107346. doi: 10.1016/j.measurement.2019.107346

Mendenhall, M. E. (2013). "Leadership and the birth of global leadership," in *Global Leadership*, 2 Edn, eds M. E. Mendenhall, J. Osland, A. Bird, G. R. Oddou, M. L. Maznevski, M. J. Stevens, et al. (New York, NY: Routledge), 1–20.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *J. Math. Psychol.* 38, 244–273. doi: 10.1006/jmps.1994.1016

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *AAAI* 6, 775–780.

Nagai, H., Yasunobu, K., Benton, C., Tsubaki, H., Takasugi, H., Shin, M., et al. (2015). *Developing Global Leadership*. Tokyo: Hakuto-Shobo Publisher.

Needham, J., and Harbsmeier, C. (1998). *Science and Civilisation in China: Volume 7, The Social Background, Part 1, Language and Logic in Traditional China*. Cambridge, MA: Cambridge University Press.

Ng, K. Y., Van Dyne, L., and Ang, S. (2009). From experience to experiential learning: cultural intelligence as a learning capability for global leader development. *Acad. Manag. Learn. Educ.* 8, 511–526.

Nimon, K., Shuck, B., and Zigarmi, D. (2015). Construct overlap between employee engagement and job satisfaction: a function of semantic equivalence? *J. Happ. Stud.* 17, 1149–1171. doi: 10.1007/s10902-015-9636-6

Norenzayan, A., Smith, E. E., Kim, B. J., and Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cogn. Sci.* 26, 653–684. doi: 10.1207/s15516709cog2605-4

Nunally, J. C., and Bernstein, I. H. (2007). *Psychometric Theory*, 3rd Edn, New York, NY: McGraw-Hill.

Osland, J. (2013). "The multidisciplinary roots of global leadership," in *Global Leadership*, 2nd Edn, eds M. E. Mendenhall, J. Osland, A. Bird, G. R. Oddou, M. L. Maznevski, M. J. Stevens, et al. (New York, NY: Routledge), 21–39.

Osland, J., Oddou, G., Bird, A., and Osland, A. (2013). Exceptional global leadership as cognitive expertise in the domain of global change. *Eur. J. Int. Manag.* 7, 517–534. doi: 10.1504/Ejim.2013.056475

Parks-Stamm, E. J., Oettingen, G., and Gollwitzer, P. M. (2010). Making sense of one's actions in an explanatory vacuum: the interpretation of nonconscious goal striving. *J. Exp. Soc. Psychol.* 46, 531–542. doi: 10.1016/j.jesp.2010.02.004

Piccolo, R. F., Bono, J. E., Heinitz, K., Rowold, J., Duehr, E., and Judge, T. A. (2012). The relative impact of complementary leader behaviors: which matter most? *Leadersh. Q.* 23, 567–581. doi: 10.1016/j.leaqua.2011.12.008

Pinker, S. (2008). *The Stuff of Though: Language As A Window Into Human Nature*. London: Penguin Books.

Podsakoff, P. M., MacKenzie, S. B., and Podsakoff, N. P. (2012). "Sources of method bias in social science research and recommendations on how to control it," in *Annual Review of Psychology*, eds S. T. Fiske, D. L. Schacter, and S. E. Taylor, (Palo Alto: Annual Reviews), 539–569.

Poli, R., Healy, M., and Kameas, A. (2010). "WordNet," in *Theory and Applications of Ontology: Computer Applications*, ed. C. Fellbaum, (New York, NY: Springer), 231–243.

Porter, M. E., and Rivkin, J. W. (2012). Choosing the United States. *Harv. Bus. Rev.* 90, 80–91.

Prinz, W., Aschersleben, G., and Koch, I. (2009). "Cognition and Action," in *Oxford Handbook of Human Action*, eds E. Morsella, J. A. Bargh, and P. M. Gollwitzer, (Boston: Oxford University Press, Inc).

Qin, Y., Li, B., and Yu, L. (2015). Management innovations in a Chinese hotel company: the case of 7 days inn. *Int. J. Contemp. Hosp. Manag.* 27, 1856–1880. doi: 10.1108/Ijchm-02-2014-0080

Renfrew, C. (1987). *Archaeology and Language. The Puzzle of Indo-European Origins*. London: Pimlico.

Russell, B. (1922). "An introduction to the tractatus logico-philosophicus," in *Tractatus Logico-Philosophicus*, ed. L. Wittgenstein, (London: Kegan Paul).

Russell, B. (1918/2007). "The relation of sense-data to physics," in *Mysticism and Logic*, ed. B. Russell, (Nottingham: Bertrand Russell Peace Foundation), 139–170.

Schaller, T. K., Patil, A., and Malhotra, N. K. (2015). Alternative techniques for assessing common method variance: an analysis of the theory of planned behavior research. *Organ. Res. Methods* 18, 177–206. doi: 10.1177/1094428114554398

Schwarz, N. (1999). Self-reports - how the questions shape the answers. *Am. Psychol.* 54, 93–105. doi: 10.1037//0003-066x.54.2.93

Semin, G. (1989). The contributon of linguistic factors to attribute inference and semantic similarity judgements. *Eur. J. Soc. Psychol.* 19, 85–100.

Sidnell, J., and Enfield, N. J. (2012). Language diversity and social action a third locus of linguistic relativity. *Curr. Anthropol.* 53, 302–333. doi: 10.1086/665697

Slaney, K. L. (2017). *Validating Psychological Constructs : Historical, Philosophical, and Practical Dimensions*. London: Palgrave MacMillan.

Slaney, K. L., and Racine, T. P. (2013). Constructing an understanding of constructs. *New Ideas Psychol.* 31, 1–3. doi: 10.1016/j.newideapsych.2011.02.010

Slobin, D. I. (1996). "From thought and language to thinking for speaking," in *Rethinking Linguistic Relativity*, eds J. Gumperz, and S. C. Levinson, (Cambridge: Cambridge University Press), 70–96.

Smedslund, J. (1988). What is measured by a psychological measure. *Scand. J. Psychol.* 29, 148–151. doi: 10.1111/j.1467-9450.1988.tb00785.x

Smedslund, J. (1994). Nonempirical and empirical components in the hypotheses of 5 social-psychological experiments. *Scand. J. Psychol.* 35, 1–15. doi: 10.1111/j.1467-9450.1994.tb00928.x

Smedslund, J. (2015). "The value of experiments in psychology," in *The Wiley Handbook of Theoretical and Philosophical Psychology*, eds J. Martin, J. Sugarman, and K. L. Slaney, (New Jersey: John Wiley & Sons, Ltd), 359–373.

Sparck Jones, K. (1986). *Synonymy and Semantic Classification*. Edinburgh, Scotland: Edinburgh University Press.

Sparck-Jones, K. (1972). A statistical interpretation of term speci- ficity and its application in retrieval. *J. Documentation*., 11–21.

Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Soc. Stud. Sci.* 48, 204–231. doi: 10.1177/0306312718772094

Tejeda, M. J., Scandura, T. A., and Pillai, R. (2001). The MLQ revisited - psychometric properties and recommendations. *Leadersh. Q.* 12, 31–52. doi: 10.1016/S1048-9843(01)00063-7

Thiesen, F. (2003). Om Språkforholdene i Pakistan I – Urdu, Pakistans Nasjonalspråk (On the language situation in Pakistan I: Urdu, the national language of Pakistan). Språknytt, (2/3).

Tsui, A. S., Nifadkar, S. S., and Ou, A. Y. (2007). Cross-national, cross-cultural organizational behavior research: advances, gaps, and recommendations. *J. Manag.* 33, 426–478. doi: 10.1177/0149206307300818

Van Dierendonck, D., Sousa, M., Gunnarsdóttir, S., Bobbio, A., Hakanen, J., Pircher Verdorfer, A., et al. (2017). The cross-cultural invariance of the servant leadership survey: a comparative study across eight countries. *Admin. Sci.* 7:8. doi: 10.3390/admsci7020008

van Knippenberg, D., and Sitkin, S. B. (2013). A critical assessment of charismatic-transformational leadership research: back to the drawing board? *Acad. Manag. Ann.* 7, 1–60. doi: 10.1080/19416520.2013.759433

van Schuur, W. H., and Kiers, H. A. L. (1994). Why factor analysis often is the incorrect model for analyzing bipolar concepts, and what models to use instead. *Appl. Psychol. Measur.* 18, 97–110.

Wang, Z. M., and Satow, T. (1994). Leadership styles and organizational effectiveness in Chinese-Japanese joint ventures. *J. Manag. Psychol.* 9, 31–48.

Warner-Søderholm, G. (2013). Beyond a literature review of Hall's context dimension: scale development, validation & empirical findings within a Norwegian study. *Int. J. Bus. Manag.* 8, 27–40.

Weseman, R. A. (2007). Review of incidence and management of chylous ascites after small bowel transplantation. *Nutr. Clin. Pract.* 22, 482–484. doi: 10.1177/0115426507022005482

Whorf, B. L. (1956). "Science and Linguistics," in *Language, Thought and Reality*, ed. J. B. Carroll, (Cambridge, MA: MIT Press).

Wu, Z., and Palmer, M. (1994). "Verbs semantics and lexical selection," in *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA.

Zander, L., Mockaitis, A. I., Harzing, A. W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., et al. (2011). Standardization and contextualization: a study of language and leadership across 17 countries. *J. World Bus.* 46, 296–304. doi: 10.1016/j.jwb.2010.07.009

Zhang, Y., Huai, M. Y., and Xie, Y. H. (2015). Paternalistic leadership and employee voice in China: a dual process model. *Leadersh. Q.* 26, 25–36. doi: 10.1016/j.leaqua.2014.01.002

Zhang, Z. Q., Gentile, A. L., and Ciravegna, F. (2013). Recent advances in methods of lexical semantic relatedness - a survey. *Nat. Lang. Eng.* 19, 411–479. doi: 10.1017/S1351324912000125

Zhang, Z. X., Chen, Z. X., Chen, Y. R., and Ang, S. (2014). Business leadership in the Chinese context: trends, findings, and implications. *Manag. Organ. Rev.* 10, 199–221. doi: 10.1111/more.12063

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 26 March 2020
doi: 10.3389/fpsyg.2020.00561

Trust and Distrust as Artifacts of Language: A Latent Semantic Approach to Studying Their *Linguistic Correlates*

David Gefen[1], Jorge E. Fresneda[2] and Kai R. Larsen[3]*

[1] Decision Sciences and MIS Department, LeBow College of Business, Drexel University, Philadelphia, PA, United States, [2] Marketing, Martin Tuchman School of Management, New Jersey Institute of Technology, Newark, NJ, United States, [3] Organizational Leadership and Information Analytics, Leeds School of Business, University of Colorado Boulder, Boulder, CO, United States

Trust and distrust are crucial aspects of human interaction that determine the nature of many organizational and business contexts. Because of socialization-borne familiarity that people feel about others, trust and distrust can influence people even when they do not know each other. Allowing that some aspects of the social knowledge that is acquired through socialization is also recorded in language through word associations, i.e., *linguistic correlates*, this study shows that known associations of trust and distrust can be extracted from an authoritative text. Moreover, the study shows that such an analysis can even allow a statistical differentiation between trust and distrust—something that survey research has found hard to do. Specifically, measurement items of trust and related constructs that were previously used in survey research along with items reflecting distrust were projected onto a semantic space created out of psychology textbooks. The resulting distance matrix of those items was analyzed by applying covariance-based structural equation modeling. The results confirmed known trust and distrust relationship patterns and allowed measurement of distrust as a distinct construct from trust. The potential of studying trust theory through text analysis is discussed.

Keywords: trust, distrust, latent semantic analysis, text analysis, machine learning, linguistic correlates

OPEN ACCESS

Edited by:
Andrea Greco, University of Bergamo, Italy

Reviewed by:
Laszlo Hunyadi, University of Debrecen, Hungary
Asyraf Afthanorhan, Sultan Zainal Abidin University, Malaysia

****Correspondence:***
Jorge E. Fresneda
fresneda@njit.edu

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

Received: *27 September 2019*
Accepted: *09 March 2020*
Published: *26 March 2020*

Citation:
Gefen D, Fresneda JE and Larsen KR (2020) Trust and Distrust as Artifacts of Language: A Latent Semantic Approach to Studying Their Linguistic Correlates. Front. Psychol. 11:561. doi: 10.3389/fpsyg.2020.00561

INTRODUCTION

Research Objective

Allowing that socialized knowledge is embedded in the language also through the tendency of words to co-occur together across relevant documents, this study argues that such *linguistic correlates* can reveal much about trust and distrust—key socialization beliefs. That proposition is supported by projecting questionnaire items about trust and distrust and their familiarity antecedent and a behavioral outcome on a semantic space (discussed below) that was built out of a relevant corpus of three psychology textbooks (Myers, 1998), and then analyzing the resulting cosine distance

matrix of those questionnaire items. The analysis shows that not only are expected theoretical correlations supported, but also that trust and distrust can be statistically differentiated in this manner—something that survey research using questionnaires had difficulty doing. The ability to mine such knowledge from language may be another tool to study human behavior through text analysis in cases where surveys cannot be given to human subjects, where the context is unknown to them, and where constructs that cannot be easily differentiated such as trust and distrust need to be studied. To clarify, we are not claiming that this method replaces surveys, only that it could complement survey research.

The Importance of Trust and Distrust in Human Behavior

Interpersonal trust is a key driver of human behavior and a key determinant of interpersonal relationships because it allows people to assume, rightly or not, that they know how those they trust will behave (Blau, 1964; Rotter, 1971; Sztompka, 1999). At the core of trust theory (Luhmann, 1979) is the recognition that people are independent agents who cannot be fully controlled and that these people are not even consistently rational in their behavior. Therefore, contends trust theory, trying to understand how others will behave can introduce so much social uncertainty as to be cognitively overwhelming to the extent that people might refrain from interacting with others they do not trust because they do not understand what is going on. Knowing how the trusted party will behave, i.e., trusting them, allows people to reduce that otherwise overwhelming social complexity to more manageable levels by assuming that the trusted party will behave in expected socially acceptable manners and not in other unexpected socially unacceptable manners (Gefen et al., 2003a).

Because it allows reducing the otherwise overwhelming social complexity to manageable levels, and in doing so allows people to assume that there is a common understanding of what behavior is permitted, interpersonal trust is a key driver of social and economic structures (Williamson, 1985; Fukuyama, 1995; Zak and Knack, 2001). Trust also determines the preference of one vendor or company over another in contracting relationships, again, presumably because the trusting party assumes it knows how the trusted party will behave (Gulati, 1995; Kumar, 1996; Gefen et al., 2008b; Greenberg et al., 2008), and whether any interaction will even occur because when the risk of not knowing what the trusted party will do is too big then people refrain from interacting (Fukuyama, 1995). Because of those reasons, trust is also a key determinant in the adoption of new IT (Gefen, 2004) of many kinds including ecommerce (Gefen et al., 2003b), virtual teams (Jarvenpaa et al., 1998), online communities (Ridings et al., 2002), online software marketplaces (Gefen and Carmel, 2008), online consumer marketplaces such as eBay (Pavlou and Gefen, 2004, 2005; Pavlou and Fygenson, 2006), e-banking (Kaabachi et al., 2017; Ofori et al., 2017), e-government (Warkentin et al., 2018), among others. Trust is even a determinant of susceptibility to phishing (Moody et al., 2017). Basically, trust is a key construct in human behavior (Schoorman et al., 2007).

Trust, as often defined in management papers, is about "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p. 712). This willingness to trust is based according to Mayer et al. (1995) on beliefs about the *trustworthiness*—ability, benevolence, and integrity—of the trusted party. That assessment of trustworthiness is modeled by Mayer et al. (1995) as the consequence of previous interactions with the trusted party. As research showed, that assessment of trustworthiness can also be the result of the trusting person's propensity to trust, often modeled as initial trust, that is based on lifelong socialization (Rotter, 1967; McKnight et al., 1998, 2002; Gefen et al., 2003b), a propensity that is influenced *inter alia* by socialization and national culture (Fukuyama, 1995). In the technology context, for ecommerce as an example, this initial trust may be even more important than the perceived usefulness and ease of use of the IT (Gefen et al., 2003a).

Distrust is closely related to trust and is an integral part of trust theory, but it is not just the opposite of trust. Even early on in the study of trust it was recognized that the breakdown of trust results in more than just a reduction in the level of trust in that such a breakdown often results in a transformation of the relationship to one of avoidance (Blau, 1964). Conceptually, distrust is a separate construct entirely from trust (Blau, 1964; Kramer, 1999; McKnight and Choudhury, 2006), dealing with negative beliefs about the other party. Although research based on survey data has found it hard to statistically differentiate between trust and distrust (Benbasat et al., 2008), neuroscience has shown that the neural correlates of trust and distrust are distinctly different (Dimoka, 2010; Riedl et al., 2010b) with trust being mostly associated with neural-correlates that are associated with rewards such as the putamen (the outer part of the lentiform nucleus of the brain) and with information processing such as the dorsolateral prefrontal cortex (DLPFC) while distrust is associated with neural correlates associated with aversion such as the insular cortex and with fear such as the amygdala. Thus, while trust brings people together based mostly on rational reasons, distrust separates them based on fear and aversion. The ability of neuroscience to identify this distinction where survey research could not do so has been one of the reasons suggested for adopting neuroscience into the mainstream of social sciences research (Dimoka et al., 2012). As this study will show, the ability of text analysis to also make this distinction is a point for consideration.

Trust, Distrust, Familiarity, and the Objective of This Study

A key reason why people trust or distrust, and the context of this study, is because people are socialized into trusting strangers (Rotter, 1971), or a specific group of strangers (Zucker, 1986), or distrusting them as the case might be (Fukuyama, 1995), through socialization and the historical and social information that that socialization conveys (Fukuyama, 1995). In a nutshell, socialization is "learned" familiarity with people at large or with

a specific group of people one has not yet encountered. This kind of learning through socialization is typically portrayed as a lifelong experience starting at childhood through education and interaction with other people. People are taught whom to trust and whom to distrust sometimes even on a purely irrational and historically and socially totally irrelevant basis as an integral part of their "education" of learned prejudices and "truisms"[1].

Across business contexts, familiarity is a significant predictor of trust. Being familiar with the trusted party means that the trusting party knows better what to expect, what the rules of conduct are, how the trusted party might react, and has a reasonable idea of the trusted party's integrity, benevolence (or at least caring), and capability based on past performance. Being familiar with the other party taps into many of the reasons why trust is needed: being able to assess the trustworthiness of the trusted party as a way of reducing risk (Mayer et al., 1995), being able to better understand what is happening and plan and respond accordingly (Luhmann, 1979, 1988; Gefen et al., 2003b), as well as reducing distrust across social group boundaries (Gefen and Ridings, 2003).

Indeed, choosing a familiar party to contract with can be so compelling an argument that often people will prefer to contract with a party they are familiar with regardless of the price (Gefen and Carmel, 2008). This is not just that trusted vendors can charge a price premium (Ba and Pavlou, 2002). It is that in some cases, specifically low cost contracts to develop software and related services, the trusted party will always win the bid over unfamiliar parties regardless of price (Gefen and Carmel, 2013). And, when price does come into play, such as in large software contracts signed by a bank, then the familiar party will on average be given the contract on terms that require less oversight such as contracting on a time and materials basis rather than a fixed price contract (Gefen et al., 2008b; Benaroch et al., 2016).

Socialization, and the familiarity it creates, is a powerful tool, but not all its teachings are direct and overt. Some of the messages that socialization broadcasts are subtle and hidden in the language we speak. Indeed, as immoral as it may be, the dictionary definition of many words, e.g., racial or social classifications, carry such social praise or stigma that make people feel that they are somewhat "familiar" with the other party based on what they were taught and thus leads them to trust or distrust total strangers based on this socialization. A rather innocuous example is the one Zucker (1986) gives of US banking in the early 1900 where people trusted bankers based on the social class of those bankers who, presumably because one was taught that they belong to a "better" social class, can be trusted. In other words, familiarity can also create distrust. The importance of familiarity in building trust, and by extension reducing distrust, seems to be true across business contexts. This applies in contracting between organizations (Williamson, 1985; Gambetta, 1988; Gulati, 1995; Bolton and Dewatripomt, 2005; Ha and Perks, 2005; Gefen et al., 2008b; Gulati and Sytch, 2008) as well as ecommerce (Gefen, 2000; Pavlou and Fygenson, 2006) and ecommerce recommendation agents (Komiak and Benbasat, 2006), as it is in daily life (Blau, 1964; Luhmann, 2000).

Accordingly, the objective of this study is to argue for linguistic socialization and its implications in a new and expanded context. We argue that *trust and distrust are registered into the very language we speak* and that therefore some aspects of the socialization into trusting and distrusting can be studied through text analysis. To emphasize this registered socialized embedded knowledge, we label it *linguistic correlates*. Technically, it is the same as analyzing how words and vectors of words correlate (or co-appear), expanding on the logic of Gefen and Larsen (2017).

The next sections will show that running text analysis on a semantic space that was built by analyzing a corpus created out of the paragraphs of three psychology textbooks (Myers, 1998)—arguably a reasonable trustworthy repository of theories on human behavior—supports this proposition. This semantic space was chosen because it is accessible in the public domain at lsa.colorado.edu together with an interface that allows projecting combinations of entire sentences on that sematic space. The result of that projection is a matrix of cosine distances that can be extracted for further analysis. That further analysis in covariance-based structural equation modeling (CBSEM) will show that projecting sentences that comprise of survey measurement items dealing with trust, distrust, and related constructs allows the reconstruction of a statistical model based on the cosine distances among each pair of those sentences. And, that in doing so, known psychological relationships of trust and of distrust can be reconstructed.

Deriving Linguistic Correlates of Trust and Distrust Through a Semantic Space

Just as the conclusions being drawn about sociological events and the interpretation of social constructs will differ based on the sources being read, so too it is recognized that the results of text analysis will depend on the corpus being analyzed and its reliability and connection to the topic being studied. Accordingly, as the study of trust and distrust is clearly in the realm of psychology, and undeniably many other social sciences related to psychology, we chose a semantic space derived from a corpus based on textbooks in psychology.

The "psychology" semantic space used in this study was created based on a total of 13,902 textbook paragraphs containing 30,119 unique terms. The approach depends on a bag-of-words representation where each paragraph's word order is abandoned and frequently used terms downweighed before the term-document matrix is subjected to a singular value decomposition (SVD) as described in Larsen and Monarchi (2004). In general practice, 300–500 dimensions are retained (Arnulf et al., 2014). In the creation of this specific semantic space a 398-dimension space was created. This means that each word that is part of one of the textbooks is represented by a 398-dimensional vector of what that term *means* in the context of all the other words. The meaning of a sentence is inferred through the addition of the vectors for each of the words in the sentence, a process known as projection. That sematic space is available in the

[1]As a demonstration of this trust-building or distrust-forming socialization process, think of how many times you heard, or maybe even gave, advisory or precautionary sociological "truisms" such as "Don't talk to strangers" or "What do you expect of (fill in your preferred racial/religious/social/political etc. noun)? They are always right/wrong/racist!."

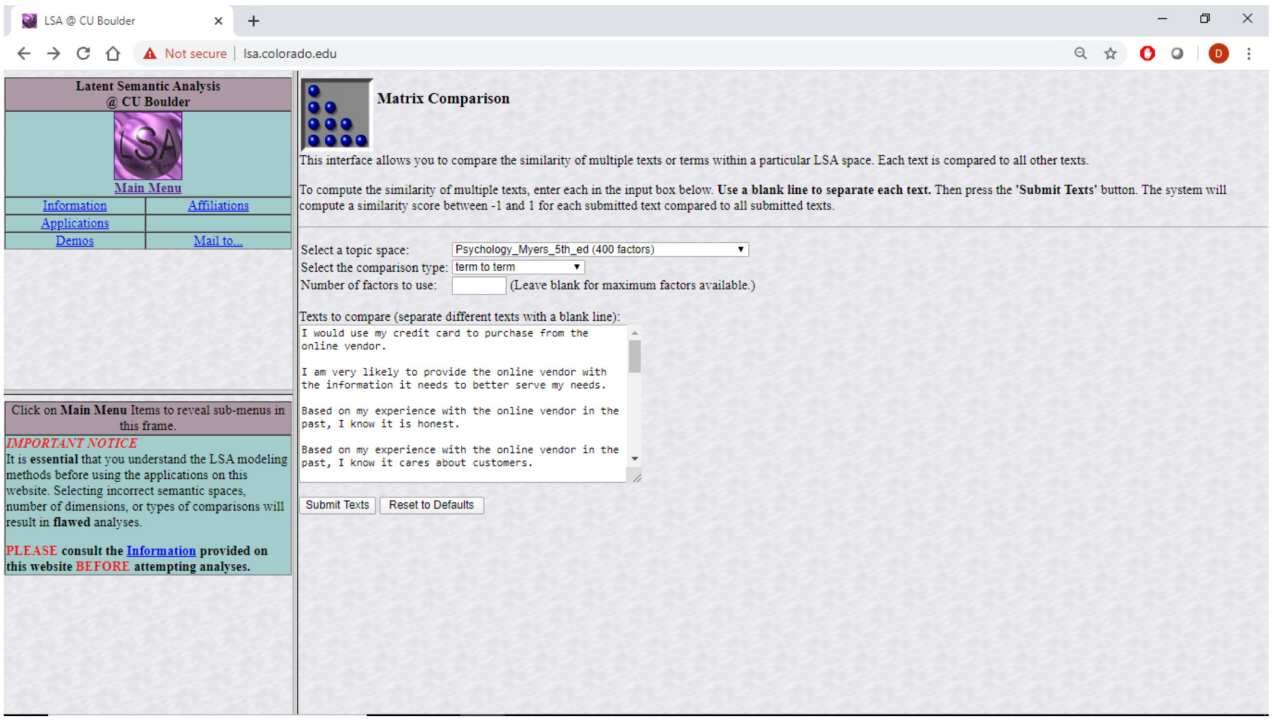


FIGURE 1 | Producing LSA correlations among questionnaire items at lsa.colorado.edu.

TABLE 1 | Measurement items semantic distance cosines produced by lsa.colorado.edu.

	USE1	USE2	TR1	TR2	TR3	TR4	TR5	TR6	TR7	DT1	DT2	DT3	FM1	FM2	FM3
USE1	1	0.8	0.83	0.83	0.82	0.82	0.83	0.83	0.82	0.76	0.83	0.83	0.76	0.78	0.76
USE2	0.8	1	0.85	0.84	0.84	0.84	0.84	0.84	0.83	0.77	0.76	0.76	0.79	0.8	0.78
TR1	0.83	0.85	1	0.98	0.99	0.97	1	1	0.97	0.79	0.77	0.77	0.8	0.82	0.8
TR2	0.83	0.84	0.98	1	0.98	0.96	0.98	0.98	0.96	0.79	0.77	0.77	0.8	0.82	0.8
TR3	0.82	0.84	0.99	0.98	1	0.97	0.99	0.99	0.97	0.82	0.76	0.75	0.8	0.82	0.8
TR4	0.82	0.84	0.97	0.96	0.97	1	0.97	0.98	0.96	0.78	0.77	0.77	0.8	0.83	0.8
TR5	0.83	0.84	1	0.98	0.99	0.97	1	1	0.97	0.79	0.77	0.76	0.8	0.82	0.8
TR6	0.83	0.84	1	0.98	0.99	0.98	1	1	0.97	0.79	0.77	0.77	0.8	0.82	0.8
TR7	0.82	0.83	0.97	0.96	0.97	0.96	0.97	0.97	1	0.77	0.76	0.75	0.79	0.82	0.79
DT1	0.76	0.77	0.79	0.79	0.82	0.78	0.79	0.79	0.77	1	0.83	0.82	0.75	0.76	0.75
DT2	0.83	0.76	0.77	0.77	0.76	0.77	0.77	0.77	0.76	0.83	1	0.98	0.78	0.78	0.78
DT3	0.83	0.76	0.77	0.77	0.75	0.77	0.76	0.77	0.75	0.82	0.98	1	0.77	0.77	0.77
FM1	0.76	0.79	0.8	0.8	0.8	0.8	0.8	0.8	0.79	0.75	0.78	0.77	1	0.92	0.91
FM2	0.78	0.8	0.82	0.82	0.82	0.83	0.82	0.82	0.82	0.76	0.78	0.77	0.92	1	0.94
FM3	0.76	0.78	0.8	0.8	0.8	0.8	0.8	0.8	0.79	0.75	0.78	0.77	0.91	0.94	1

public domain through an interface at lsa.colorado.edu, shown in **Figure 1**.

Specifically, survey items from previous research that dealt with trust were projected into this semantic space together with items dealing directly with distrust. The cosine distances among the projected survey items as produced by lsa.colorado.edu were then analyzed using CBSEM. The results discussed in the next sections are as theory predicts. Specifically, the questionnaire items were copied into lsa.colorado.edu, shown in **Figure 1**, and the derived cosine distances, shown in **Figure 2**, were then copied and arranged in a matrix form ready to be analyzed with Mplus, shown in **Table 1**. The questionnaire items appear in **Table 2**.

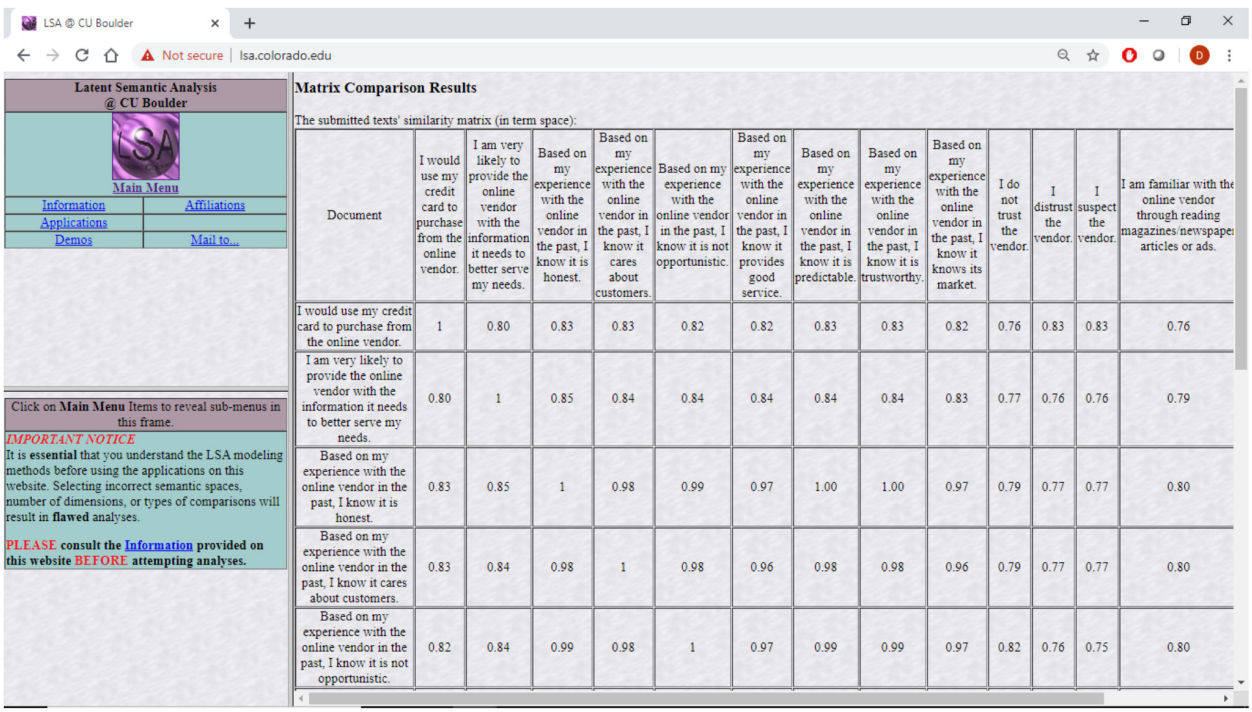
The Potential of Studying Linguistic Correlates in the Study of Trust and Distrust

Showing, as this study does, that studying the word associations of trust and distrust produces equivalent results as survey research on trust did, raises the possibility, and clearly more

TABLE 2 | Measurement items projected on the Myers (1998) textbook semantic space.

Code	Construct/measurement items	Standardized loading
	Intended Use	
USE1	I would use my credit card to purchase from the online vendor	0.90***
USE2	I am very likely to provide the online vendor with the information it needs to better serve my needs	0.89***
	Trust	
TR1	Based on my experience with the online vendor in the past, I know it is honest	0.99***
TR2	Based on my experience with the online vendor in the past, I know it cares about customers	0.99***
TR3	Based on my experience with the online vendor in the past, I know it is not opportunistic	Dropped
TR4	Based on my experience with the online vendor in the past, I know it provides good service	0.98***
TR5	Based on my experience with the online vendor in the past, I know it is predictable	Dropped
TR6	Based on my experience with the online vendor in the past, I know it is trustworthy	Dropped
TR7	Based on my experience with the online vendor in the past, I know it knows its market	0.98***
	Distrust	
DT1	I do not trust the vendor	0.84***
DT2	I distrust the vendor	0.99***
DT3	I suspect the vendor	0.99***
	Familiarity with the e-Vendor	
FM1	I am familiar with the online vendor through reading magazines/newspaper articles or ads	0.95***
FM2	I am familiar with the online vendor through visiting the site and searching for CDs/books	0.98***
FM3	I am familiar with the online vendor through purchasing CDs/books at this site	0.96***

****means significant at the 0.001 level.*



Document	I would use my credit card to purchase from the online vendor.	I am very likely to provide the online vendor with the information it needs to better serve my needs.	Based on my experience with the online vendor in the past, I know it is honest.	Based on my experience with the online vendor in the past, I know it cares about customers.	Based on my experience with the online vendor in the past, I know it is not opportunistic.	Based on my experience with the online vendor in the past, I know it provides good service.	Based on my experience with the online vendor in the past, I know it is predictable.	Based on my experience with the online vendor in the past, I know it is trustworthy.	Based on my experience with the online vendor in the past, I know it knows its market.	I do not trust the vendor.	I distrust the vendor.	I suspect the vendor.	I am familiar with the online vendor through reading magazines/newspaper articles or ads.
I would use my credit card to purchase from the online vendor.	1	0.80	0.83	0.83	0.82	0.82	0.83	0.83	0.82	0.76	0.83	0.83	0.76
I am very likely to provide the online vendor with the information it needs to better serve my needs.	0.80	1	0.85	0.84	0.84	0.84	0.84	0.84	0.83	0.77	0.76	0.76	0.79
Based on my experience with the online vendor in the past, I know it is honest.	0.83	0.85	1	0.98	0.99	0.97	1.00	1.00	0.97	0.79	0.77	0.77	0.80
Based on my experience with the online vendor in the past, I know it cares about customers.	0.83	0.84	0.98	1	0.98	0.96	0.98	0.98	0.96	0.79	0.77	0.77	0.80
Based on my experience with the online vendor in the past, I know it is not opportunistic.	0.82	0.84	0.99	0.98	1	0.97	0.99	0.99	0.97	0.82	0.76	0.75	0.80

FIGURE 2 | Resulting semantic distances of LSA correlations among questionnaire items at lsa.colorado.edu.

research is needed before such an argument can be made unequivocally, that studying the linguistic registration of trusting behavior in an appropriate source (a textbook on human psychology in this case) might allow new avenues for studying trust and distrust. Such avenues might allow the studying of trust and distrust also in contexts that cannot be studied or do not exist anymore. The context might have changed and the people not available anymore, but at least their study as they are registered linguistically can still be done. This might include studies such as how the meaning and importance of trust

and distrust as registered through word associations changed overtime. Given that one cannot administer questionnaires to people who lived in London 150 years ago, but one has easy access to the books written by Charles Dickens and others of that period, such a possibility might open the door to new understandings.

Such an approach to studying trust and distrust—and by extension other constructs, beliefs, attitudes, behaviors, etc.—might also reveal, in a broader context, why non-native speakers of English answer the same questions differently in English versus in their native language, even when the surveys are an exact translation of each other (Harzing, 2005). This approach might potentially also point out possible reasons for social differences about trust and distrust, and provide support for the hypothesized effect of history on trust and distrust as portrayed by Fukuyama (1995). Indeed, comparing the word associations of trust and distrust and the meaning revealed through those in the books of Charles Dickens compared to Henrik Ibsen might be quite revealing.

Moreover, and perhaps this is going on a tangent, if indeed part of our socialization as humans is registered in the language we speak through word correlations, then this might be especially important in predicting how people might understand the role trust and distrust play also in as of yet not quite there technologies. To put this into perspective, research on how we as people trust and distrust others has been about another party that is *human* or composed of a group of people. Specifically, in that past research the trusted party may have been a person [e.g., Blau (1964)], a community [e.g., Ridings et al. (2002)], a market populated by people [e.g., Pavlou and Gefen (2004)], an organization [e.g., Mayer et al. (1995)], a government [e.g., Warkentin et al. (2018)], or a human-like IT interface such as an avatar (Bente et al., 2008; Keeling et al., 2010). But what about a trusted party whose intentions and intelligence are not human or related to people?

Being able to understand, even if only through the knowledge embedded in language, why people trust or distrust in such a case may prove essential with the growing influx of AI into daily lives where AI is creating an environment that is sometimes beyond human understanding, as demonstrated recently in a case of a self-taught AI beating the world champion in *gos* without the world champion even understanding some of the strategies the AI applied (Economist, 2017). The linguistic correlates of trust and distrust might enable modeling human reaction also in such cases of interacting with an AI where the reasons cited above for the importance of trust and distrust do not readily apply. After all, there are no rational assessments of the behavior of an AI agent playing *go*, nor are there considerations of risk, familiarity, social strata considerations, social identification, etc. Nonetheless, being able to model in statistical terms the human response to such a world could be revealing.

The next sections will describe the method we applied to study the linguistic correlates of trust and distrust, why theoretically one might expect there to be linguistic correlates, and some details about the method, and then report the statistical analysis and discuss the results and their potential.

MATERIALS AND METHODS

Replicating the established hypotheses that familiarity builds trust, and adding to it that familiarity may also lead to the opposite, i.e., distrust, as Fukuyama (1995) relates, and further extending into both trust and distrust as major considerations in the decision to purchase online (Gefen, 2000; Dimoka, 2010), the research model is presented in **Figure 3**. This figure shows the output of the standardized Mplus analysis on the model. Boxes represent the measurement items, which in this case are the questionnaire items that were projected onto the semantic space. These items and their codes appear in **Table 2**. The covariance among all pairs of those measurement items is constrained in CBSEM so that only the covariance values associated with the paths that are shown in the model as arrows are expressed. All other covariance values are fixed at zero. Fixing those paths to zero frees enough degrees of freedom to include in the model also latent variables, i.e., constructs that while they cannot be measured directly are reflected by the explicit measurement items, as well as how those constructs relate to each other. In this formalization, each measurement item is a function of the latent variable it is assigned to, the circles, and of an error term. For example, fm1, being one of the familiarity measurement items, is predicted by the construct "familiarity" with a path estimate of 0.946 and standard error of 0.006 as well as by a random error term with a path estimate of 0.106 and a standard error of 0.012. The model of the paths leading to the measurement items is known as the measurement model. The paths among the latent variables is known as the structural model. The structural model is what the theory talks about. For example, that trust affects use is shown by the path between the circle labeled trust and the circle labeled use. Those latter paths represent the underlying proposition that the pattern of findings, i.e., supported hypotheses, as revealed in previous survey and archival data research methods can be extracted through linguistic correlates derived from an appropriate corpus.

Preparing the Model for Study

The model was tested by projecting the [Intended] Use, Trust, and Familiarity scales based on Gefen et al. (2003b) and *ad hoc* items of Distrust on the *psychology* semantic space at lsa.colorado.edu. These questionnaire items are shown in **Table 2** with the subsequent Mplus estimated standardized loadings of each item on its related latent variable (construct). The first column contains the item code. This code appears also in **Table 1** and in the Mplus code in the **Appendix**. The second column shows the wording of each item, with a header to make it easier to identify which items relate to which construct. The third column contains the standardized loading of that item on the latent variable, i.e., construct, as produced by the Mplus analysis.

The lsa.colorado.edu site receives as input a set of sentences (or individual words) that are to be projected onto one of several preexisting semantic spaces. See **Figure 1**. It then builds the cosine distances matrix of each sentence from each other sentence by running a latent semantic analysis (LSA) process. See **Figure 2**. The process involves projecting each possible pair of sentences as two vectors, each comprising all the words in one of the

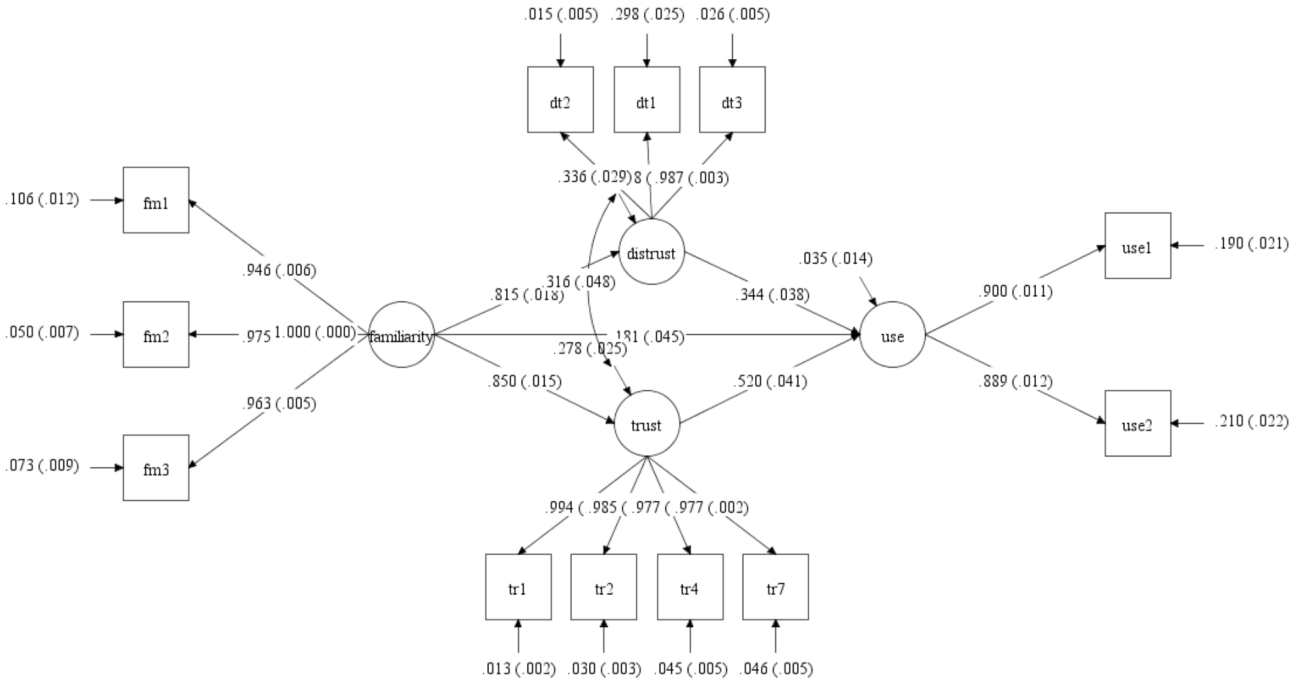


FIGURE 3 | Research model and standardized estimates produced by Mplus.

sentences, on a chosen preexisting sematic space. The idea behind LSA is that words ("terms" in LSA parlance) that tend to appear together have shared dimensions of meaning.

What LSA does is to first create a term to document [frequency] matrix (TDM) of the original corpus, possibly preparing the data beforehand through stemming and other methods, weighing the terms, and then applying SVD to the TDM to reduce the dimensionality of the data (Dumais et al., 1988; Deerwester et al., 1990). It is then assumed that words that appear together on the same principal component (dimension) after this dimensionality reduction exercise share some meaning (Landauer and Dumais, 1997; Landauer et al., 1998). Words can appear in many principal components thus showing the richness of language and that the same word can carry many meanings. The result of the SVD is known as a semantic space. The semantic space analyzed already exists on the lsa.colorado.edu site. The vectors of the sentences can then be projected onto this sematic space, even though the sentences themselves never existed in the original texts. The comparison of these vectors allows a calculation of the cosine distance between them.

At its core, LSA is about word co-occurrences. It is a data-driven approach, and some therefore see it as more objective (Evangelopoulos et al., 2012). As argued, certain words tend to be used together, such as "trust" and "purchase," so words take on meanings both in terms of the words with which they co-occur, and in terms of words with which they do not co-occur frequently, such as "sky" and "purchase." Words that co-occur frequently will tend to have a smaller cosine distance between them, and, by extension, two sentences where each contains words that tend to appear in the other sentence will also have a small cosine distance between them. Importantly, LSA works in cases of second and third-level relationships where words do not even need to co-occur, but both co-occur with the same words. For example, LSA will tend to recognize that terms such as "distrust" and "trust" are related even if the words never co-occurred in the text analyzed, for example because both may appear together with the word "transaction" or the word "relationship."

Because these co-occurrences reflect language used to describe the world, the LSA word vectors contain within them reflections of our shared perceptions of how the world works. Much work has gone into understanding how LSA works relative to the human mind, and Landauer (2007, p. 31) even argued that LSA "demonstrates a computational method by which a major component of language learning and use can be achieved." The applicability of LSA to partially replicate through text analysis survey responses by people seems to support this contention (e.g., Arnulf et al., 2014, 2018; Gefen and Larsen, 2017). Without entering the debate of what LSA does or does not do [cf., for example, Valle-Lisboa and Mizraji (2007)], we use LSA to address a specific question in a way that is mathematically rigorous and that can be replicated by anyone with an understanding of statistical methods.

More details on how to run LSA in R together with a discussion of the methodological and statistical validity consideration are available at Gefen et al. (2017). As LSA is now widely accepted as a research method, with hundreds of uses within Psychology and Information Systems, we will not go into further depth on the process. Readers interested in this

process are referred to one of many detailed descriptions, ranging from mathematical introductions (e.g., Larsen and Monarchi, 2004; Martin and Berry, 2007) to conceptual explanations (e.g. Evangelopoulos et al., 2012; Arnulf et al., 2014).

We chose LSA for several reasons. First, it is an established and tested method and has been so for the last two decades (Tonta and Darvish, 2010; Evangelopoulos et al., 2012). Second, it has been shown to simulate human thought processes, producing survey results that sometimes correspond to how human subjects answer the same questionnaire items (Larsen et al., 2008; Arnulf et al., 2014; Gefen and Larsen, 2017), including assessing the meaning of words through their association with other words (Yeari and van den Broek, 2014; Bhatia, 2017), and even simulating priming effects through word choice (Günther et al., 2016). LSA has even been applied in this context to support the supposition that the meaning of a word is derived through its associations to other words (Kintsch and Mangalath, 2011), and supporting that supposition even by comparing the LSA semantic meaning of a word with eye tracking (Huettig et al., 2006). And, third, the method we apply, running a CBSEM analysis on the correlations derived from LSA semantic spaces has been previously applied to show that the widely supported model of IT adoption, the technology acceptance model (TAM) (Davis, 1989), can be supported by projecting the existing scales of that model on a semantic space that was created out of unrelated newspaper articles (Gefen and Larsen, 2017).

The Underlying Idea Behind Linguistic Correlates

As specified, the idea being propagated in this study is that socialization knowledge is to some extent ingrained in the language that we speak and write. And that this applies also in word co-occurrence relationships. As a result of this engraining, analyzing word co-occurrence relationships in relevant text could reveal some of that socialization knowledge. Such an argument is supported by the significant and consistent replication of the relationships between the perceived usefulness and the perceived ease of use scales of TAM (Davis, 1989) in both the measurement model (how items load significantly only on their assigned constructs and not on other constructs) and the correlation between the constructs in the structural model by projecting its questionnaire items on two newspaper semantic spaces (Gefen and Larsen, 2017).

The argument for ingrained knowledge in language, expanding on the proposition advanced by Gefen and Larsen (2017), is that if certain words or combinations of words tend to occur together, then these co-occurrence tendencies might be registering socialized knowledge linguistically. Thus, for example, if the word "distrust" and the word "avoid" tend to occur together considerably more than "trust" and "avoid" do, while "trust" tends to co-occur often with "purchase" than "distrust" does then this co-occurrence might be registering that people tend to avoid that which they distrust but tend to purchase from those they trust.

This kind of analysis may actually have the potential to reveal self-censored knowledge too, addressing a known problem with questionnaires. It is well-known that people completing surveys, even anonymous ones, consider both what they think the survey administer wants to hear and what they themselves are implying by their answers (Cook and Campbell, 1979; Shadish et al., 2002). Thus, it would be rather hard to elicit honest non-politically correct prejudices because people completing a questionnaire know that expressing such ideas openly is shunned by society, meaning that there is a bias in such data if it is collected through surveys. However, because LSA analyzes also indirect associations among words, it might catch such prejudices. Indeed, indirect associations of terms identified by LSA has been shown to be beneficial in the case of analyzing medical records to reveal important patterns in the population being studied (Gefen et al., 2018) as well as how IT design battles evolve in the press (Miller et al., 2018). Moreover, terms that are not easily distinguished from each other in the statistical analysis of survey questionnaire items filled by people, might nonetheless be differentiated in text analysis because they each have their own distinct associations with other terms. This differentiation will indeed be shown in the next section.

This is not an argument for causation. It does not mean that people behave as they do because of that linguistically ingrained knowledge, as implied in the "Sapir-Whorf hypothesis" (Hill and Mannheim, 1992) that language determines thoughts and behavior or in an Orwellian control of thought through a *newspeak* language (Orwell, 1948). Rather, the argument is for correlation. People behave as they do for a myriad of reasons, and the language they and others use reflects those tendencies. It may be that their behavior—and more accurately in this case their story-telling about their behavior—reflects their socialization through language, but it may just as well be that language registers the shared aspects of theirs's and many others' story-telling.

ANALYSIS RESULTS

Analysis Process

The measurement items' cosine matrix produced by lsa.colorado.edu was entered as input to Mplus version 7.4 and analyzed as a reflective CBSEM. In our measurement model, the reflective CBSEM measurement items are modeled as reflecting a latent variable, known otherwise as a construct. Thus, DT1, DT2, and DT3 all reflect the latent variable (construct) Distrust, and no other construct, while USE1 and USE2 reflect the latent variable Use, and no other, etc. If there are significant cross-loadings, i.e., a loading of a measurement item on a construct it was not assigned to, then CBSEM will identify that cross-loading in the modification index table together with an assumed χ^2 improvement as well as a noticeable change in the overall fit indices of the model. The measurement model part of a CBSEM model specifies that pattern of measurement items to constructs loadings. The structural model then specifies the relationship among those constructs. Mplus analyzes both the measurement model and the structural model together, highlighting any problems with unspecified covariance or with measurement items whose covariance overlaps. It is standard procedure in CBSEM to drop items that have such

problems (SAS, 2013), but it should be reported (Gefen et al., 2011) as we do here.

Items TR5 and TR6 were dropped because the cosine distance between them and between each of them and TR4 was 1.000, meaning that as far as the maximum likelihood algorithm that CBSEM applies as a default for continuous variables these three items are practically indistinguishable from each other. Being indistinguishable from each other, results mathematically in an Mplus observation that "the sample covariance matrix could not be inverted" when those items were included. No other pairs of measurement items had a cosine of 1.000 between them. Item TR3 was dropped to improve model fit (including TR3 did not change the overall model pattern but resulted in an RMSEA of 0.138). It is long established as an acceptable practice to drop items in CBSEM because of such reasons (Bollen, 1989; Jöreskog and Sörbom, 1989).

The Mplus analysis was run specifying that the sample size was 400, which is the rounded number of dimensions created by lsa.colorado.edu for the textbooks when creating the semantic space. As is standard in Mplus for continuous measurement items, we retained the default maximum likelihood analysis. Overall model fit was acceptable (Gefen et al., 2011): χ^2_{48} = 187.853, RMSEA = 0.085, CFI = 0.985, TLI = 0.979. The Mplus code is available in the **Appendix**.

Interpretation of the Analysis

The standardized structural model showed that Use was significantly predicted by Trust ($\beta = 0.52$, $p < 0.001$), Distrust ($\beta = 0.18$, $p < 0.001$), and Familiarity ($\Gamma = 0.34$, $p < 0.001$).[2] That Trust is a stronger predictor of Use than Familiarity is consistent with anthropological studies where knowing the historical context determines levels of trust and distrust that, in turn, determine behavioral intentions [e.g., Fukuyama (1995)]. These significant predictors of Use are consistent with the literature cited above. Familiarity significantly predicted Trust ($\Gamma = 0.79$, $p < 0.001$) and Distrust ($\Gamma = 0.82$, $p < 0.001$). This too is consistent with the literature cited above.

The CBSEM model modeled Trust and Distrust as being correlated on account of these two constructs being portrayed in theory as non-overlapping opposite beliefs/assessments of each other with non-overlapping opposite consequences on behavioral intentions (Blau, 1964; Luhmann, 1979; Sztompka, 1999). The theoretical distinction between the Trust and Distrust constructs is also supported by fMRI studies (Dimoka, 2010; Riedl et al., 2010b). The distinction between Trust and Distrust as separate constructs is supported in the CBSEM model through the very low modification index values among the items of the Trust and Distrust constructs. Trust and Distrust as constructs are significantly correlated ($\theta = 0.32$, $p < 0.001$).

R^2 values were 0.97 for Use, 0.72 for Trust, and 0.66 for Distrust. Cross-loadings were low, as also indicated through the acceptable levels of the RMSEA statistic. Notice that LSA does not specify the sign (plus or minus) of the cosine distances. Hence, the Mplus model shows that the relationships between Distrust and all the other constructs are positive. That is a known limitation of LSA in that it measures the semantic closeness of words, or vectors of words such as the entire sentences of a questionnaire item, as an angle but where the direction of that angle is immaterial.

Ad Hoc Analysis

As an additional *ad hoc* analysis to establish that differentiating between Trust and Distrust indeed produces a significantly better model, a model that unites these two constructs was compared with the original model. Specifically, the χ^2 of the original model (χ^2_{48} = 187.853) was compared with the χ^2 of an alternative model in which Trust and Distrust were united into one construct. The resulting χ^2 of this alternative model (χ^2_{51} = 1073.722) was significantly worse ($\Delta\chi^2_3$ = 855.869), showing that separating Trust and Distrust produces a significantly better model.

DISCUSSION

Summary of the Results

The proposition advanced in this study was that socialized knowledge is also ingrained in language, and that this registered knowledge can be extracted through text analysis tools such as LSA and subsequent statistical analysis. These linguistic correlates, as we call them, can be analyzed to both reconstruct existing hypotheses, and do so purely through text analysis and without resorting to distributing surveys to human subjects, as well as be applied to additional analyses not easily performed through survey research. This proposition was demonstrated in the context of studying trust and distrust as they relate to familiarity as an antecedent and to purchase (labeled "use" in other studies) as an outcome.

The analysis supports this proposition, but also highlights some text analysis nuances that should be considered. The analysis shows that linguistic correlates can be analyzed to support the measurement model, showing that the cosine distances between pairs of questionnaire items that are projected on a relevant semantic space can then be analyzed through CBSEM to support the expected significant loadings of those questionnaire items on the latent variable they theoretically reflect. The linguistic correlates also enabled the statistical differentiation between trust and distrust (see *ad hoc* analysis in section "Interpretation of the Analysis"), which has been hard to do with survey research (Gefen et al., 2008a) even though this distinction is suggested in theory (e.g., Fukuyama, 1995; Blau, 1964) and has been shown in neural science (e.g., Dimoka, 2010; Riedl et al., 2010b). The analysis also supports the next part of the proposition that the correlation patterns among those constructs, i.e., the structural model, are consistent with theory. The analysis,

[2]In CBSEM notation, exogenous (independent) variables (latent constructs) are labeled β, while endogenous (dependent) variables are labeled ξ. Paths between exogenous and endogenous variables are labeled Γ. Paths between endogenous variables are labeled β. Correlation paths between endogenous variables are labeled θ. Thus, because Familiarity is modeled as affecting Trust, Distrust, and Use, all the paths leading out of Familiarity are labeled Γ. The paths between Trust and Use and between Distrust and Use are labeled β, and the correlation between Trust and Distrust is θ.

however, also shows that the cosine distance between some pairs of items was 1.000, i.e., a perfect overlap, producing a result that is seldom seen in data collected through surveys administered to human subjects, and requiring dropping items accordingly.

The conclusion is that some aspects of socialized knowledge about trust and distrust are ingrained in the language we speak, and that that the registration of this socialized knowledge can be extracted through linguistic correlates to the extent that allows recreating relationships that theory implies.

Implications for Trust Theory and the Possible Role of Linguistic Correlates

Trust theory and the English language clearly differentiate between trust and distrust, showing that although the two terms are related in their contexts, they are not the same and do not even overlap in their meaning. Such a difference is shown also in this study where both trust and distrust are correlated to familiarity and to use as well as to each other, but their items significantly do not reflect the same, one, latent construct. That studying linguistic correlates could show that difference when survey research that analyzes human subjects' responses to questionnaires could not, and thereby possibly creating a misinterpretation that trust and distrust overlap in meaning, shows a potential contribution for analyzing linguistic correlates, or at least that linguistic correlates can add significantly to knowledge acquired through survey research.

More specifically from a trust theory perspective, that Trust had a stronger standardized effect on Use ($\beta = 0.52$, $p < 0.001$) than Familiarity ($\Gamma = 0.18$, $p < 0.001$) did, suggests that, as previous models [e.g., Gefen (2000)] show, it is mainly that familiarity builds trust and that it is mostly trust rather than familiarity that determines behavior. Extending that line of logic, that the standardized effect of Trust is considerably stronger than that of Distrust ($\beta = 0.34$, $p < 0.001$) suggests that trust is more important in determining behavior than distrust is in the context of providing information online (see wording of the USE1 and USE2 items) as projected on this specific semantic space. Likewise, that Familiarity affects both Trust ($\Gamma = 0.85$, $p < 0.001$) and Distrust ($\Gamma = 0.82$, $p < 0.001$) with an almost equal standardized coefficient and that those coefficients are considerably higher than the standardized correlation between Trust and Distrust ($\theta = 0.32$, $p < 0.001$), suggests that familiarity affects trust and distrust through two mostly unrelated channels. Such an observation is consistent with how Fukuyama (1995) describes the evolution of trust and of distrust in different cultures differently based on their histories. What builds trust is not what creates distrust.

Such an ability to differentiate between trust and distrust was brought a decade ago by the burgeoning NeuroIS discipline. (NeuroIS is a name given to the discipline and society that studies neuroscience as applied to information systems). NeuroIS used that same need to differentiate between trust and distrust (e.g., Dimoka, 2010; Riedl et al., 2010b). NeuroIS then used that verification of the trust-distrust distinction through neural correlates to argue that because neuroscience could do so while questionnaire data research could not, to advance a key argument for the importance of such neuroscience research (Riedl et al., 2010a; Dimoka et al., 2012). The same argument may be applicable to text analysis and to linguistic correlates too. Not only can the study of linguistic correlates support behavioral hypotheses through the patterns of word co-occurrences, but it can even support hypotheses that survey data may not be able to. Neuroscience and text analysis are clearly not the same and they undeniably measure different data. Nonetheless, building on that same argument about the ability to study if two constructs might not be the same even when survey research cannot show it, text analysis does have the advantage over neuroscience in that it is cheaper and faster. There are potentially many other such constructs of interest that could be studied.

Broader Implications for Text Analysis in View of Linguistic Correlates

As Gefen and Larsen (2017) previously suggested, analyzing linguistic correlates may also add another tool to the toolbox that social scientists apply to assess, and maybe statistically control for, priming (Cook and Campbell, 1979), and the inevitable introduction of common method variance in data collected by surveys (Podsakoff et al., 2003; Malhotra et al., 2006). Moreover, text analysis, even if its results do not fully overlap survey analysis given to live subjects, may also provide a cheaper option to pretest existing questionnaires before embarking on a more costly data collection endeavor with subjects. To that, this study adds also the ability to statistically show the discriminant validity, i.e., to differentiate, between constructs that theoretically and linguistically are not the same, but that survey research has not been able to show their discriminant validity.

Moreover, this kind of a method might be especially applicable to the study of contexts that cannot be studied by surveys, such as those unrelated to current actual experiences. Studying linguistic correlates might allow a glimpse into how people in the past thought, and, hence, how concepts of interest changed in their linguistic meaning and associations over time. Clearly talking to actual people or studying actual responses to surveys has its advantages, but there is no known current technology that allows us to ask Charles Dickens or Henrik Ibsen about their take on trust. Studying their writings is an obvious alternative. This method allows doing so semi-automatically. Likewise, such a method could allow studying how these linguistic correlates changed over time by comparing current literature with that of the past.

The comparison of linguistic correlates might also reveal hints as to why, as the Introduction brought, non-native speakers of English answer the same questions differently in English compared to answering the surveys in their native languages, even when the surveys are an exact translation of each other

(Harzing, 2005). It may well be that part of the answer is that the linguistic correlates of the constructs being studied in those surveys differ across languages.

Studying linguistic correlates might also reveal partially how people in the present might respond to technologies of the future. That is, studying linguistic correlates could provide a partial picture of the socialized knowledge embedded in the language aspect of why people do what they do. It might be impossible to study how people will react to new technologies such as new aspects of AI that are not available yet—and why in the context of this study they may trust or distrust those—but, looking into people's linguistic correlates might reveal at least the socialized knowledge embedded language aspect of that question. It might also reveal some hints as to why some cultures might be more open than others to accepting and trusting such AI. Such a glimpse could be of much importance considering that current theories about trust are geared at a person, group of people, or an anthropomorphized party. Current theories of trust address such a target by discussing reasons such as controlling risk and understanding the social environment. It is questionable if and how any of those reasons might apply to an AI. Studying linguistic correlates might at least identify possible motivations and drives that are socialized into language. This also suggests an avenue for possible future research into why people might trust or distrust even when the reasons provided by current research, such as controlling risk (Mayer et al., 1995) or simplifying the social environment to manageable levels (Luhmann, 1979; Gefen et al., 2003b), clearly do not apply. Possibly, such a study of trust and distrust through language usage patterns as revealed through text analysis of a reasonably expert source such as textbooks may allow assessing how people might trust and distrust also in contexts that are beyond their ability to assess risks in or to understand.

Limitations

The study demonstrated the linguistic correlates proposition through an admittedly simple model. But the very fact that the model could be replicated at all suggests that indeed at least some aspects of social knowledge are recorded in language through the association of words. Presumably, as discussed above, this ingrained knowledge corresponds to how people think either because they learned or socialized that language embedded knowledge or because that language embedded knowledge recorded how people behave. Obviously, replication with other relevant corpora is necessary, but that the analysis supported the proposition is revealing.

Limitations that apply to CBSEM would apply to this method too. Had the model been too complex then the "noise" of covariances that are not included in the model would eventually result in overall poor fit indices. Likewise, many of the overall fit indices, such as χ^2 and RMSEA are negatively affected as the sample size increases. As the tendency in LSA is to have about 300 to 500 dimensions, and therefore the analysis would be modeled as a sample size of between 300 and 500 data points, the risk of having overall fit indices that do not match the criteria we apply to survey research may become an issue.

Likewise, as with other types of data collection, it is imperative that the source of data be a reliable, valid, and relevant one. This applies in this context much as it does to interviewing experts or giving out surveys. Choosing the correct population (or corpus in this case) is crucial.

Possibly, the limitation that most limits this study and others like it is that the semantic distance, a cosine distance in this case, signifies the strength of the relationship but not its direction, i.e., whether the relationship is positive or negative. Thus, the path from Distrust to Use is positive while according to theory it should be negative. The current method does not address this. Refinements are needed to add a sign value to the cosine values produced by LSA or any other text analysis method that is applied to extract semantic distances.

CONCLUSION

This study demonstrated the ability to apply LSA and CBSEM combined to investigate the linguistic correlates of trust and distrust. The study also showed that analyzing linguistic correlates can be applied to differentiate between trust and distrust—something survey research had difficulty in doing. Clearly, the concept of linguistic correlates and the potential of modeling their role in human decision making, is not limited to trust and distrust alone. Nor is this potential limited to the study of only the present. Texts of the past could be just as readily analyzed in the method demonstrated in this paper, opening through linguistic correlates a view to the past and how people in long gone periods might have thought. Practically, this also opens the window to the possible study of how we as people of the present might respond to future technologies and contexts based on our current linguistic correlates.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

DG initiated the idea, built the theory section, ran the CBSEM analyses, and led the project and write-up. JF contributed to the LSA interpretation and to the discussion and write-up. KL ran the LSA analyses, and contributed the discussion and LSA write-up.

FUNDING

This work was supported by the Martin Tuchman School of Management at the New Jersey Institute of Technology.

REFERENCES

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Bong, C. H. (2014). Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour. *PLoS One* 9:e106361. doi: 10.1371/journal.pone.0106361

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Egeland, T. (2018). The failing measurement of attitudes: How semantic determinants of individual survey responses come to replace measures of attitude strength. *Behav. Res. Methods* 50, 2345–2365. doi: 10.3758/s13428-017-0999-y

Ba, S., and Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: price premiums and buyer behavior. *MIS Q.* 26, 243–268.

Benaroch, M., Lichtenstein, Y., and Fink, L. (2016). Contract design choices and the balance of ex-ante and ex-post transaction costs in software development outsourcing. *MIS Q.* 40, 57–82. doi: 10.25300/misq/2016/40.1.03

Benbasat, I., Gefen, D., and Pavlou, P. A. (2008). Introduction to the JMIS special issue trust in online environments. *JMIS* 25, 5–12.

Bente, G., Rüggenberg, S., Krämer, N. C., and Eschenburg, F. (2008). Avatar-mediated networking: increasing social presence and interpersonal trust in net-based collaborations. *Hum. Commun. Res.* 34, 287–318. doi: 10.1111/j.1468-2958.2008.00322.x

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychol. Rev.* 124, 1–20. doi: 10.1037/rev0000047

Blau, P. M. (1964). *Exchange and Power in Social Life*. New York, NY: Wiley.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: John Wiley and Sons.

Bolton, P., and Dewatripomt, M. (2005). *Contract Theory*. Cambridge, MA: The MIT Press.

Cook, T. D., and Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Q.* 13, 319–340.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 391–407.

Dimoka, A. (2010). What Does the Brain tell us about Trust and Distrust? Evidence from a functional neuroimaging study. *MIS Q.* 34, 373–396.

Dimoka, A., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Gefen, D., et al. (2012). On the Use of Neurophysiological Tools in IS Research: Developing a Research Agenda for NeuroIS. *MIS Q.* 36, 679–702.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). "Using latent semantic analysis to improve access to textual information," in *Paper Presented at the Conference on Human Factors in Computing Systems: Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, Washington, DC.

*Economist, T. (2017). *Artificial Intelligence. The Latest AI Can Work Things Out without Being Taught*. Available online at: https://www.economist.com/science-and-technology/2017/10/21/the-latest-ai-can-work-things-out-without-being-taught (accessed November 10, 2017).

Evangelopoulos, N., Zhang, X., and Prybutok, V. R. (2012). Latent semantic analysis: five methodological recommendations. *Eur. J. Inf. Syst.* 21, 70–86. doi: 10.1057/ejis.2010.61

Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. New York, NY: The Free Press.

Gambetta, D. (1988). "Can we Trust Trust?" in *Trust: Making and Breaking Cooperative Relations*, ed. D. Gambetta (New York, NY: Blackwell), 213–237.

Gefen, D. (2000). E-commerce: the role of familiarity and trust. *Omega* 28, 725–737. doi: 10.1016/s0305-0483(00)00021-9

Gefen, D. (2004). What makes ERP implementation relationships worthwhile: linking trust mechanisms and ERP usefulness. *J. Manag. Inf. Syst.* 23, 263–288. doi: 10.1080/07421222.2004.11045792

Gefen, D., Benbasat, I., and Pavlou, P. A. (2008a). A research agenda for trust in online environments. *J. Manag. Inf. Syst.* 24, 275–286. doi: 10.2753/mis0742-1222240411

Gefen, D., and Carmel, E. (2008). Is the World Really Flat? A look at offshoring in an online programming marketplace. *MIS Q.* 32, 367–384.

Gefen, D., and Carmel, E. (2013). Why the first provider takes it all: the consequences of a low trust culture on pricing and ratings in online sourcing markets. *Eur. J. Inf. Syst.* 22, 604–618. doi: 10.1057/ejis.2012.49

Gefen, D., Endicott, J., Fresneda, J., Miller, J., and Larsen, K. R. (2017). A guide to text analysis with latent semantic analysis in r with annotated code studying online reviews and the stack exchange community. *Commun. Assoc. Inf. Syst.* 41, 450–496. doi: 10.17705/1cais.04121

Gefen, D., Karahanna, E., and Straub, D. W. (2003a). Inexperience and experience with online stores: the importance of TAM and trust. *IEEE Trans. Eng. Manag.* 50, 307–321. doi: 10.1109/tem.2003.817277

Gefen, D., Karahanna, E., and Straub, D. W. (2003b). Trust and TAM in online shopping: an integrated model. *MIS Q.* 27, 51–90.

Gefen, D., and Larsen, K. R. (2017). Controlling for lexical closeness in survey research: a demonstration on the technology acceptance model. *J. Assoc. Inf. Syst.* 18, 727–757. doi: 10.17705/1jais.00469

Gefen, D., Miller, J., Armstrong, J. K., Cornelius, F. H., Robertson, N., Smith-McLallen, A., et al. (2018). Identifying patterns in medical records through latent semantic analysis. *Commun. ACM* 61, 72–77. doi: 10.1145/3209086

Gefen, D., and Ridings, C. (2003). IT acceptance: managing user - IT group boundaries. *ACM SIGMIS Database* 34, 25–40. doi: 10.1145/937742.937746

Gefen, D., Rigdon, E., and Straub, D. W. (2011). An update and extension to SEM guidelines for administrative and social science research. *MIS Q.* 35, III–XIV.

Gefen, D., Wyss, S., and Lichtenstein, Y. (2008b). Business familiarity as risk mitigation in software development outsourcing contracts. *MIS Q.* 32, 531–551.

Greenberg, P. S., Greenberg, R. H., and Antonucci, Y. L. (2008). The role of trust in the governance of business process outsourcing relationships A transaction cost economics approach. *Bus. Process Manag. J.* 14, 593–608. doi: 10.1108/14637150810903011

Gulati, R. (1995). Does Familiarity Breed Trust? The Implications of Repeated Ties for Contractual Choice in Alliences. *Acad. Manag. J.* 38, 85–112. doi: 10.5465/256729

Gulati, R., and Sytch, M. (2008). Does familiarity breed trust? Revisiting the antecedents of trust. *Manage. Decis. Econ.* 29, 165–190. doi: 10.1002/mde.1396

Günther, F., Dudschig, C., and Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: evidence from priming studies. *Q. J. Exp. Psychol.* 69, 626–653. doi: 10.1080/17470218.2015.1038280

Ha, H.-Y., and Perks, H. (2005). Effects of consumer perceptions of brand experience on the web: brand familiarity, satisfaction and brand trust. *J. Consum. Behav.* 4, 438–452. doi: 10.1002/cb.29

Harzing, A.-W. (2005). Does the use of English-language questionnaires in cross-national research obscure national differences? *Int. J. Cross Cult. Manag.* 5, 213–224. doi: 10.1177/1470595805054494

Hill, J. H., and Mannheim, B. (1992). Language and World View. *Annu. Rev. Anthropol.* 21, 381–404.

Huettig, F., Quinlan, P. T., McDonald, S. A., and Altmann, G. T. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychol.* 121, 65–80. doi: 10.1016/j.actpsy.2005.06.002

Jarvenpaa, S. L., Knoll, K., and Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *J. Manag. Inf. Syst.* 14, 29–64. doi: 10.1080/07421222.1998.11518185

Jöreskog, K. G., and Sörbom, D. (1989). *LISREL7: A Guide to the Program and Applications*, 2nd Edn. Chicago, IL: SPSS Inc.

Kaabachi, S., Ben Mrad, S., and Petrescu, M. (2017). Consumer initial trust toward internet-only banks in France. *Int. J. Bank Mark.* 35, 903–924. doi: 10.1108/ijbm-09-2016-0140

Keeling, K., McGoldrick, P., and Beatty, S. (2010). Avatars as salespeople: communication style, trust, and intentions. *J. Bus. Res.* 63, 793–800. doi: 10.1016/j.jbusres.2008.12.015

Kintsch, W., and Mangalath, P. (2011). The construction of meaning. *Top. Cogn. Sci.* 3, 346–370. doi: 10.1111/j.1756-8765.2010.01107.x

Komiak, S. Y. X., and Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Q.* 30, 941–960.

Kramer, R. M. (1999). Trust and distrust in organizations: emerging perspectives, enduring questions. *Annu. Rev. Psychol.* 50, 984–993.

Kumar, N. (1996). The power of trust in manufacturer-retailer relationships. *Harv. Bus. Rev.* 74, 92–106.

Landauer, T. K. (2007). "LSA as a theory of meaning," in *Handbook of Latent Semantic Analysis*, eds T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (Mahwah, NJ: Larence Erlbaum Associates, Publishers), 3–34.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295x.104.2.211

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Process.* 25, 259–284.

Larsen, K. R., and Monarchi, D. E. (2004). A mathematical approach to categorization and labeling of qualitative data: the latent categorization method. *Sociol. Methodol.* 34, 349–392. doi: 10.1111/j.0081-1750.2004.00156.x

Larsen, K. R., Nevo, D., and Rich, E. (2008). Exploring the semantic validity of questionnaire scales. *Paper Presented at the Hawaii International Conference on System Sciences*, Waikoloa, HI.

Luhmann, N. (1979). *Trust and Power*, ed. T. F. German (trans.). Brisbane: John Wiley and Sons.

Luhmann, N. (1988). "Trust: making and breaking cooperative relations," in *Familiarity, Confidence, Trust: Problems and Alternatives*, ed. D. Gambetta (Oxford: Basil Blackwell), 94–107.

Luhmann, N. (2000). Familiarity, confidence, trust: problems and alternatives," in *Trust: Making and Breaking Cooperative Relations*, ed. D. Gambetta (Oxford: University of Oxford), 94–107.

Malhotra, N. K., Kim, S. S., and Patil, A. (2006). Common method variance in is research: a comparison of alternative approaches and a reanalysis of past research. *Manag. Sci.* 52, 1865–1883. doi: 10.1287/mnsc.1060.0597

Martin, D. I., and Berry, M. W. (2007). "Mathematical foundations behind latent semantic analysis," in *Handbook of Latent Semantic Analysis*, eds D. S. M. Thomas, K. Landauer, S. Dennis, and W. K. Mahwah (Mahwah, NJ: Lawrence Erlbaum Associates).

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Acad. Manag. Rev.* 20, 709–734. doi: 10.5465/amr.1995.9508080335

McKnight, D. H., and Choudhury, V. (2006). "Distrust and trust in B2C E-commerce: Do they differ?," in *Proceedings of the International Conference on Electronic Commerce*, Fredericton.

McKnight, D. H., Choudhury, V., and Kacmar, C. (2002). Developing and validating trust measures for e-commerce: an integrative typology. *Inf. Syst. Res.* 13, 334–359. doi: 10.1287/isre.13.3.334.81

McKnight, D. H., Cummings, L. L., and Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Acad. Manag. Rev.* 23, 473–490. doi: 10.5465/amr.1998.926622

Miller, J., Narayanan, V. K., Gefen, D., and Larsen, K. R. (2018). "Exploration of a design battle using latent semantic analysis," in *Paper presented at the Academy of Management Proceedings*, Chicago, IL.

Moody, G. D., Galletta, D. F., and Dunn, B. K. (2017). Which phish get caught? An exploratory study of individuals' susceptibility to phishing. *Eur. J. Inf. Syst.* 26, 564–584. doi: 10.1057/s41303-017-0058-x

Myers, D. G. (1998). *Psychology,* 5th Edn. New York, NY: Worth Publishers.

Ofori, K. S., Boateng, H., Okoe, A. F., and Gvozdanovic, I. (2017). Examining customers' continuance intentions towards internet banking usage. *Mark. Intell. Plan.* 35, 756–773. doi: 10.1108/mip-11-2016-0214

Orwell, G. (1948). *1984*. Chennai: Rupa.

Pavlou, P. A., and Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: an extension of the theory of planned behavior. *MIS Q.* 30, 115–143.

Pavlou, P. A., and Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Inf. Syst. Res.* 15, 37–59. doi: 10.1287/isre.1040.0015

Pavlou, P. A., and Gefen, D. (2005). Psychological contract violation in online marketplaces: antecedents, consequences, and moderating role. *Inf. Syst. Res.* 16, 372–399. doi: 10.1287/isre.1050.0065

Podsakoff, P. M., Lee, J. Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *J. Appl. Psychol.* 88, 879–903. doi: 10.1037/0021-9010.88.5.879

Ridings, C., Gefen, D., and Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *J. Strateg. Inf. Syst.* 11, 271–295. doi: 10.1016/s0963-8687(02)00021-5

Riedl, R., Banker, R. D., Benbasat, I., Davis, F. D., Dennis, A. R., Dimoka, A., et al. (2010a). On the foundations of NeuroIS: reflections on the Gmunden Retreat 2009. *Commun. Assoc. Inf. Syst.* 27, 243–264.

Riedl, R., Hubert, M., and Kenning, P. (2010b). Are there neural gender differences in online trust? An fMRI study on the trustworthiness of eBay offers. *MIS Q.* 34, 397–428.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *J. Pers.* 35, 651–665. doi: 10.1111/j.1467-6494.1967.tb01454.x

Rotter, J. B. (1971). Generalized Expectancies for Interpersonal Trust. *Am. Psychol.* 26, 443–450.

SAS (2013). *SAS/STAT® 13.1 User's Guide The CALIS Procedure*. Cary, NC: SAS Institute Inc.

Schoorman, F. D., Mayer, R. C., and Davis, J. H. (2007). An integrative model of organizational trust: past, present, and future. *Acad. Manag. Rev.* 32, 344–354. doi: 10.5465/amr.2007.24348410

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Sztompka, P. (1999). *Trust: A Sociological Theory*. Cambridge: Cambridge University Press.

Tonta, Y., and Darvish, H. R. (2010). Diffusion of latent semantic analysis as a research tool: a social network analysis approach. *Journal of Informetr.* 4, 166–174. doi: 10.1016/j.joi.2009.11.003

Valle-Lisboa, J. C., and Mizraji, E. (2007). The uncovering of hidden structures by latent semantic analysis. *Inf. Sci.* 177, 4122–4147. doi: 10.1016/j.ins.2007.04.007

Warkentin, M., Sharma, S., Gefen, D., Rose, G. M., and Pavlou, P. (2018). Social identity and trust in internet-based voting adoption. *Gov. Inf. Q.* 35, 195–209. doi: 10.1016/j.giq.2018.03.007

Williamson, O. E. (1985). *The Economic Institutions of Capitalism*. New York, NY: The Free Press.

Yeari, M., and van den Broek, P. (2014). The role of textual semantic constraints in knowledge based inference generation during reading comprehension: a computational approach. *Memory* 23, 1193–1214. doi: 10.1080/09658211.2014.968169

Zak, P. J., and Knack, S. (2001). Trust and growth. *Econ. J.* 111, 295–321. doi: 10.1111/1468-0297.00609

Zucker, L. G. (1986). "Production of trust: institutional sources of economic structure, 1840-1920," in *Research in Organizational Behavior*, Vol. 8, eds B. M. Staw and L. L. Cummings (Greenwich, CN: JAI Press), 53–111.

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

APPENDIX

Mplus Code

```
TITLE:      Familiarity to Trust and Distrust to Use based on Psychology textbook
DATA:       FILE IS t2.txt;
            Type=fullcorr;
            Nobs=400;
VARIABLE:   NAMES ARE
  Use1 Use2
  TR1-TR7
  FM1-FM3
  DT1 DT2 DT3;

  usev Use1 Use2
  TR1 TR2 TR4 TR7
  FM1 FM2 FM3
  DT1 DT2 DT3;

ANALYSIS:   Estimator=ML;

MODEL:      Familiarity BY FM1-FM3;
            Distrust By DT1 DT2 DT3;
            Trust BY TR1 TR2 TR4 TR7;
            Use BY Use1 Use2;

            USE on Trust Familiarity Distrust;
            Trust Distrust on Familiarity;
            Trust with Distrust;
```

Output: SAMPSTAT modindices stdyx Tech4 CROSSTABS RESIDUAL.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 21 April 2020
doi: 10.3389/fpsyg.2020.00526

The Effect of User Psychology on the Content of Social Media Posts: Originality and Transitions Matter

Lucia Lushi Chen, Walid Magdy and Maria K. Wolters*

School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

Multiple studies suggest that frequencies of affective words in social media text are associated with the user's personality and mental health. In this study, we re-examine these associations by looking at the transition patterns of affect. We analyzed the content originality and affect polarity of 4,086 posts from 70 adult Facebook users contributed over 2 months. We studied posting behavior, including silent periods when the user does not post any content. Our results show that more extroverted participants tend to post positive content continuously and that more agreeable participants tend to avoid posting negative content. We also observe that participants with stronger depression symptoms posted more non-original content. We recommend that transitions of affect pattern derived from social media text and content originality should be considered in further studies on mental health, personality, and social media.

Keywords: affect, social media, emotion, Facebook, personality traits, depression, mental health, non-original content

OPEN ACCESS

Edited by:
Jan Ketil Arnulf,
BI Norwegian Business School, Norway

Reviewed by:
David Gallardo-Pujol, University of Barcelona, Spain
Michael R. Sciandra, Fairfield University, United States

****Correspondence:***
Lucia Lushi Chen
lushi.chen@ed.ac.uk

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

Received: *30 November 2019*
Accepted: *05 March 2020*
Published: *21 April 2020*

Citation:
Chen LL, Magdy W and Wolters MK (2020) The Effect of User Psychology on the Content of Social Media Posts: Originality and Transitions Matter. Front. Psychol. 11:526. doi: 10.3389/fpsyg.2020.00526

1. INTRODUCTION

Many people express rich moods and emotions in their social media posts. Psychologists use the word "affect" to describe these experiences of feelings and emotions. Affect plays an important role in cognition (Gross et al., 1998) and well-being (Silvera et al., 2008). Therefore, affective expressions in social media text have emerged as a key variable for making inferences about users' personality traits (Golbeck et al., 2011; Bachrach et al., 2012; Farnadi et al., 2013) or mental health (De Choudhury et al., 2013; Coppersmith et al., 2014; De Choudhury and De, 2014; Bazarova et al., 2015).

Existing studies formulate the associations between affect and well-being based on the frequencies of affective words used in social media text (Yarkoni, 2010; Golbeck et al., 2011; Schwartz et al., 2013; Park et al., 2015; Chen et al., 2020). However, patterns of affect are an important class of symptoms of affective disorders (Frijda, 1993; Rottenberg, 2005; Bylsma et al., 2011; Carlo et al., 2012; Thompson et al., 2012; Houben et al., 2015; Sheppes et al., 2015). Personality may also predispose individuals to specific moods (Rusting and Larsen, 1995; Rusting, 1998). With this in mind, we examined how patterns of affect expressed in social media text are related to users' mental health and personality.

While non-original content has been extensively studied in opinion mining (Balahur et al., 2009; Agarwal et al., 2011), it has been comparatively neglected in the study of psychological interpretations of social media data. However, social media users often use lyrics or quotes to communicate their emotions. Such content comes from other media, such as literature, videos, films, or music, which can evoke strong emotional experiences (Scherer and Zentner, 2001; Juslin and Laukka, 2004; Scherer, 2004). Since the affect of the non-original content may be different from

the social media users' affect when they are post this content, we differentiated between original and non-original content in our analysis.

This pilot study was designed to examine the following research questions:

1. **Changes in Affect:** To what extent do changes in the affect of social media posts correlate with users' personality traits and mental well-being?
2. **Originality:** To what extent does the use of non-original material in their posts correlate with users' personality traits and mental well-being?

Following best practice in sentiment analysis and opinion mining, we distinguish between positive, negative, neutral, and mixed (both positive and negative) affect (Moilanen and Pulman, 2007; Agarwal et al., 2011; Rosenthal et al., 2015).

We used a well-known dataset, myPersonality (Bachrach et al., 2012; Youyou et al., 2015), which enriches Facebook posts with many validated psychological measures. In MyPersonality, positive mental well-being is measured using the Satisfaction with Life Scale (Diener et al., 1985, 1999), while the presence of depressive symptoms is assessed using the Centre for Epidemiologic Studies Depression scale (CES-D) (Radloff, 1977). Personality traits are established following the OCEAN model (McCrae and John, 1992), which consists of the five traits Openness to Experience, Conscientiousness, Extroversion, Agreeableness, and Neuroticism.

We included all 70 adult users who provided sufficient, regular Facebook data for 2 months before completion of the CES-D questionnaire and corrected for multiple comparisons in our statistical analysis. We find that the transitions from one affective state to another expressed in social media posts give us a highly nuanced view of personality traits. While the amount of non-original posts in ones' social media status updates is closely linked to depression symptoms, this link is mediated by neuroticism.

2. BACKGROUND

Affect refers to both mood and emotion. Moods are slow-moving states that can be influenced by people, objects or situations, whereas emotions are quick reactions to stimuli (Watson, 2000; Rottenberg and Gross, 2003) and are highly situation- or object-specific (Bylsma et al., 2008). Mood influences the probability of having emotions of the same valence—negative mood facilitates negative emotions, and positive mood makes positive emotions more likely (Fredrickson, 1998; Rottenberg, 2005). Affect is an important predictor of mental well-being, including a person's overall satisfaction with life (Headey et al., 1993; Singh and Jha, 2008; Chen et al., 2017), and the level of symptoms of depression (Coppersmith et al., 2015; Resnik et al., 2015; Tsugawa et al., 2015).

Personality also predisposes people to certain affective states (Rothbart et al., 2000). While neuroticism is associated with negative affect (Pishva et al., 2011), positive affect is strongly linked to extroversion (Fujita et al., 1991; Watson and Clark, 1997). Extroverts experience more positive affect because they engage in more social situations (Diener and Emmons, 1984; Ryan and Deci, 2001). Individuals who score high on agreeableness have a greater ability to regulate negative affect (Meier et al., 2006; Haas et al., 2007). This relationship between affect and personality is also reflected in social media studies (Pennebaker and King, 1999; Golbeck et al., 2011; Schwartz et al., 2013; Lin et al., 2017). For example, people who use negative affective words in their social media posts tend to have lower conscientiousness, lower agreeableness (Golbeck et al., 2011), and higher neuroticism (Pennebaker and King, 1999).

In psychology, quantitative representations of affect are typically multidimensional (Russell, 1980). In this study, we focus on valence, which is represented in many classic affect models. Traditional measures, such as the Positive and Negative Affect Schedule (PANAS) (Watson et al., 1988), report the strength of positive and negative valence. Mixed valence can occur when people experience "dialectic" emotion, which is a mix of positive and negative emotions (Schimmack et al., 2002; Russell, 2003).

The personality trait measurements in myPersonality are based on Costa and McCrae's well-validated OCEAN model (McCrae and John, 1992). The model consists of five dimensions: extroversion, agreeableness, conscientiousness, neuroticism, and openness to experience. Neuroticism refers to the degree of emotional stability. Openness reflects the degree of creativity and curiosity. Conscientious individuals tend to be careful and diligent. Extroversion refers to a tendency to be energetic and friendly. Agreeableness reflects the tendency to be compassionate and to cooperate with others (Digman, 1990). The five-factor structure has proved to be robust in both self and peer ratings (McCrae and John, 1992), in both children and adult (Mervielde et al., 1995), and across different cultures (McCrae and Allik, 2002) and to be stable over time (McCrae and John, 1992).

3. DATA AND METHODOLOGY

The myPersonality data set (Bachrach et al., 2012; Youyou et al., 2015) contains more than 180,000 Facebook users, enriched with a variety of additional validated scales (Bachrach et al., 2012). The collection of myPersonality data complied with the terms of service of Facebook, informed consent for research use was obtained from all users, and researchers had to seek permission to use the dataset. Permission for the use of this database was obtained before it closed for new studies in 2018. The study was granted Ethical Approval by the Ethics Committee of the School of Informatics, University of Edinburgh.

3.1. Choice of Scales

From the extensive data collected within myPersonality, we chose two scales for quantifying mental well-being, the *Center for Epidemiologic Studies Depression Scale (CES-D)* and the *Satisfaction with Life Scale (SWL)*. The CES-D scale measures a key aspect of mental health, the presence of depression symptoms (Radloff, 1977). The scale has high internal consistency, test-retest reliability (Radloff, 1977; Roberts, 1980; Orme et al., 1986), and validity (Orme et al., 1986). Following

previous social media studies (Park et al., 2012; De Choudhury et al., 2013), we adopt a score of 22 or higher as a cut-off value for likely depressive disorder (maximum score: 60). The five-item SWL scale has been tested across different cultures and age groups (Pavot and Diener, 2009) and has been found to have high internal consistency and temporal reliability (Diener et al., 1985). Personality traits were measured using a 100-item scale using items from the open-source International Personality Item Pool (Goldberg et al., 2006) that were validated against the NEO-PI-R (Schwartz et al., 2013) instrument.

3.2. Selection of Participants

The data set was originally designed for a study of the effect of mental well-being and values on social media disclosure. We therefore selected only those participants who had completed the CES-D scale, the SWL scale, and the Schwartz Value survey (Schwartz, 1992) in addition to the full personality questionnaire. A total of 301 participants in myPersonality provided full data for all four scales.

To ensure we had enough posts to assess the frequency of affect transitions, we only included users in our sample that regularly updated their public Facebook feed (*regular users*). We defined regular users as individuals who posted on average twice a week or more. We estimated posting frequency using the average post-count per day during the sampling frame. If an individual had a post-count per day of 0.3, this individual made around 110 posts in 365 days, which was roughly equivalent to an average of two posts per week. Of the original 301 participants, 122 (40.5%) were regular users.

Since the CES-D asks about symptoms in the past week, we excluded a further 31 users who had not posted any content in the week before completing the CES-D scale. We then focused on a 60-days span (2 months) before CES-D completion to ensure that we had sufficient data to track the development of users' moods. We removed 14 users who contributed <20 posts during that time. Finally, we removed four users who were under 18 years old and three users with more than 20% of the posts written in a language other than English, because English was the common language of the annotation team. The final sample consisted of 4,086 posts from 70 users.

3.3. Corpus Annotation

3.3.1. Social Media Affect

For the purpose of this study, we refer to the affect shown in social media posts as *social media affect*. In this study, following (Mohammad, 2016), we operationalize valence as the post-author's attitude toward a primary target of opinion. We refer to the "dialectic" affective state as *mixed valence*. If there is no clear trend toward positive or negative affect, the associated valence is *neutral*.

After extensive piloting, we created an annotation guideline (available as part of the supplementary material) that was largely based on Mohammad (2016)'s work on defining the valence of a social media post. Each post is assigned one of four affect polarities: + (positive), − (negative), ± (mixed), or 0 (neutral). We used manual annotation since this is commonly used in computational linguistics to create a baseline gold standard data set for further analysis (Teufel, 1999).

Of the 4,086 posts, 2,698 (66%) were annotated by a team of six trained annotators and 1,185 (29%) by the first author; 5% of all posts were annotated by all seven annotators to establish inter-rater reliability, which was measured using Cohen's κ (Gamer et al., 2019). Average inter-rater reliability between the first author and the annotators is 0.88, and it is 0.78 among the six annotators.

After annotation, most of the posts were of positive valence (N = 1,588, 39%), followed by negative valence (N = 1,164, 28%), neutral valence (N = 982, 24%), and mixed valence (N = 312, 8%). A total of 40 posts were excluded from analysis because they did not contain English text.

3.3.2. Originality

We define posts that consist of quotes from sources, such as song lyrics, books, or movies as non-original content; all other content was defined as original. Since non-original content might not directly reflect the user's moods or emotions, annotators were instructed to annotate such posts according to the likely emotions of the author. For example, if a post consists of an uplifting motivational quote, annotators considered the underlying valence to be positive.

In order to establish the originality of a post, we retrieved the first page of results obtained by searching for the post-text using the Google API. For each web page on the first page of results, we computed the cosine similarity between the post-content and the page content. Posts with a cosine similarity >0.96 were labeled as non-original, and posts with a cosine similarity between 0.92 and 0.96, where the website links or website names included the words "lyrics" or "quote" were labeled as potentially non-original. Posts with a cosine similarity lower than 0.92 were labeled as original. The cutoff points were determined based on a sample of 300 posts manually annotated for originality by the first author. On these posts, the classifier yields 100% recall, 81% precision, and an F1-score of 0.89. In our data set, 287 (7%) of all posts were identified as non-original.

3.4. Modeling Affect Transitions

We examine two types of transitions:

- **Post-level vs. Day-level:** *Post-level* transitions focus on changes in affect between subsequent social media posts, whereas *day-level* transitions focus on changes in overall dominant affect between subsequent days.
- **Silence vs. Non-silence:** Not all users post every day. In our *default* models, these silent days are ignored, whereas in our *with-silence* models, days without posts are explicitly modeled as *Silence*.

The post-level social media affect is likely to be influenced by *underlying emotions*, which change more quickly, whereas the day-level social media affect is likely to be influenced by *underlying mood* during the day. Day-level affect was calculated as follows. If the majority of the posts p_{ij} on day d_j have the same affect a, then the affect of day d_j is set to a. If there is an equal number of positive (+) and negative (−) posts or if the number

TABLE 1 | Affect and originality representation for a sample week.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Affect							
Post-level	+ −	+ − +	+ +	S	±	0	+ −
Day-level	−	+	+	S	±	0	±
Originality							
Post-level	O N O	O O N	N N	S	O	O	N N

↔; −, negative valence; +, positive valence; ±, mixed valence; S, silence day; O, original content; N, non-original content.

of mixed affect (±) posts is equal to the number of posts with other types of affect, affect is set to ± (mixed). For transitions between original and non-original posts, we only consider the post-level representation. **Table 1** shows an example of the affect and originality representations.

3.5. Statistical Analysis

Demographic differences between users above and below the CES-D cut-off score for probable depression were assessed using Wilcoxon-Mann-Whitney tests (R-package "Stats").

We used Pearson correlation coefficients to assess the significance of correlations between social media data on the one hand and personality traits and mental well-being on the other hand. Due to the small sample size and the number of correlations computed, all correlation coefficients were estimated using a permutation approach (Higgins, 2003), as implemented in the R Package `jmuOutlier` (Garren, 2017). Correlations that reach $p < 0.01$ or better are reported as significant; correlations that reach $p < 0.05$ are reported as trends in the data. For all correlations reported in the paper, we give the estimated correlation coefficient, the bootstrap 95% confidence interval, and the corresponding coefficient of determination r^2.

4. RESULTS

4.1. Demographics and Baseline Statistics

Table 2 shows the basic statistics of our sample. Our data predominantly comes from single female Caucasian young adults. The average CES-D score is above the cut-off for possible depressive disorder.

Thirty-nine (56%) participants had a CES-D score of 22 or higher (mean: 33, SD: 6.5), which means that it is possible that they have depressive disorder, and 31 (44%) had a score of 21 or lower (mean: 12, SD: 6).

Participants with possible depressive disorder are less extroverted ($Z = 375$, $p < 0.005$) and have higher levels of neuroticism ($Z = 990$, $p < 0.001$), lower levels of conscientiousness ($Z = 375$, $p < 0.001$), and lower satisfaction with life ($Z = 323$, $p < 0.001$). Detailed results are reported in **Figure 1** Plot 1.

All scales are normally distributed (Shapiro-Wilks test), except for openness to experience ($W = 0.96$, $p < 0.05$) and satisfaction with life ($W = 0.95$, $p < 0.05$), which are bimodal. **Figure 1** Plot 1 shows the correlations between different personality dimensions. As expected, the five personality dimensions are not orthogonal.

TABLE 2 | Demographics of the sample.

Variable	*N* (%)	Variable	Mean (SD)
Gender		**Age**	
- Female	49 (70%)	- Female	23.52 (6.56)
- Male	21 (30%)	- Male	22.84 (7.13)
Ethnicity		**Personality**	
- Caucasian	54 (75%)	- Openness to Experience	4.19 (0.46)
- Black	3 (4%)	- Conscientiousness	3.20 (0.75)
- Asian	5 (7%)	- Extraversion	3.11 (3.83)
- Other	8 (14%)	- Agreeableness	3.55 (0.68)
		- Neuroticism	2.98 (0.89)
Living status		**Mental well-being**	
- Living with partner	8 (10%)	- SWL	4.18 (1.44)
- Single	54 (77%)	- CES-D	23.79 (11.86)
- Married	5 (7%)		
- Unknown	3 (4%)		

Caucasian includes White people of American, British, and other origins; Black includes African-Americans and Black people from Europe. SWL, score for Satisfaction with Life Scale; CES-D, Center for Epidemiologic Studies Depression Scale.

4.2. Social Media Affect: Frequencies vs. Transitions

For **overall frequencies of affect category**, the only clear correlation is between extroversion and positive content. Overall, more extroverted participants are more likely to have days where they make predominantly positive posts ($r = 0.29$, $p < 0.01$, 95%CI = (−0.15, 0.32), $r^2 = 0.08$). In addition, participants who score higher on agreeableness tend to post fewer negative posts and have fewer days with predominantly negative posts [both $r = -0.26$, $p < 0.05$, 95%CI = (−0.48, −0.04), $r^2 = 0.07$].

When we look at **transitions between affect categories**, however, a more nuanced picture emerges. **Table 3** summarizes the correlations between personality, well-being and transition types. Significant correlations are summarized in **Table 4**. Due to the number of correlations presented, we choose a cut-off of $p < 0.01$, which is stricter than the normal $p < 0.05$.

Several transition types are correlated positively and negatively with Extroversion and Agreeableness. Neuroticism, conscientiousness, and SWL show interesting trends ($p < 0.05$) that do not reach significance (c.f. **Table 3**).

More extroverted participants are more likely to post predominantly positive content several days in a row [*day-level*, +↔+, $r = 0.30$, $p < 0.001$, 95% CI = (0.06, 0.54), $r^2 = 0.09$]. They have more transitions to or from a silence day with a positive post [*post-level with-silence*, S↔+, $r = 0.29$, $p < 0.01$, 95% CI = (−0.01, 0.46), $r^2 = 0.08$]. This pattern fits well with the overall predominance of posts with positive affect. Extroverts are also less likely to alternate between days with neutral and days with non-neutral content [*day-level*, for both 0↔+ and 0↔−, $r = -0.28$, $p < 0.01$, 95% CI = (−0.52, −0.09), $r^2 = 0.08$].

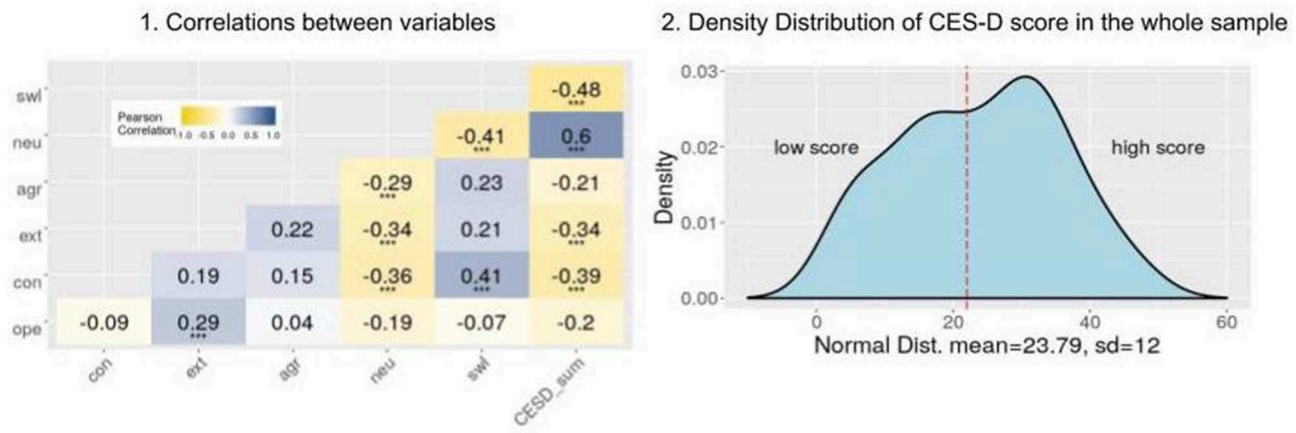


FIGURE 1 | Basic statistics for personality trait scores, SWL and CES-D scores. Plot 1 is a heat map of correlations between personality traits, SWL, and CES-D scores (***$p < 0.001$). Plot 2 illustrates the distribution of the CES-score in the entire sample ($N = 70$). The dotted line indicates the cutoff score of 22.

People who score higher on agreeableness are less likely to follow a post with negative affect with another negative-affect post [−↔−, *post-level with-silence*: $r = -0.37$, $p < 0.001$, 95% CI = (−0.50, −0.06), $r^2 = 0.14$]. This tendency is much less pronounced on the day level [−↔−, $r = -0.22$, $p < 0.1$, 95% CI = (−0.44, −0.02), $r^2 = 0.04$]. On top of that, they are more likely to alternate between days with mixed valence and silence [*day-level*, ±↔S, $r = 0.28$, $p < 0.01$, 95% CI = (−0.01, 0.46), $r^2 = 0.08$, post-level with-silence, ±↔S, $r = 0.29$, $p < 0.01$, 95% CI = (0.08, 0.52), $r^2 = 0.08$].

Participants with higher neuroticism tend to alternate between positive and negative content, but this is only evident when we take silence into account [+↔−, *post-level with-silence*: $r = 0.23$, $p < 0.05$, 95% CI = (0.00, 0.47), $r^2 = 0.04$, *post-level without-silence*: $r = 0.16$, 95% CI = (−0.08, 0.41), $r^2 = 0.025$, *day-level*: $r = 0.21$, $p < 0.05$, 95% CI = (−0.46, −0.10), r^2=0.04].

There are interesting differences in transition patterns that incorporate information about silence days and those that do not. When disregarding silence days, we observe that people with higher conscientiousness or extroversion are slightly less likely to follow a neutral post with another neutral post [*post-level without-silence*, conscientiousness, 0↔0, $r = -0.23$, $p < 0.05$, 95% CI = (−0.41, −0.04), $r^2 = 0.07$; extroversion, 0↔0, $r = -0.24$, $p < 0.05$, 95% CI = (−0.41, −0.04), $r^2 = 0.07$].

When we take into account silence days for computing transitions, we find several more interesting trends. People who are more satisfied with life are more likely to follow a neutral post with another neutral post [0↔0, *day-level*: $r = 0.25$, $p < 0.05$, 95% CI = (−0.01, 0.44), $r^2 = 0.06$]. In addition, people with higher neuroticism are more likely to alternate between positive and negative posts [0↔−, *day-level*: $r = 0.21$, $p < 0.05$, 95% CI = (−0.01, 0.40), $r^2 = 0.04$] but less likely to make a positive post after one or more silence days [S↔+, *post-level with-silence*: $r = -0.22$, $p < 0.05$, 95% CI = (−0.48, 0.00), $r^2 = 0.04$]. We found that silence-to-silence transitions are not correlated with personality or mental health.

4.3. Post-originality

High CES-D scores are significantly correlated with posting non-original content [$r = 0.29$, $p < 0.01$, 95% CI = (0.10, 0.46), $r^2 = 0.08$]. There is a similar tendency for participants with higher neuroticism scores [$r = 0.25$, $p < 0.05$, 95% CI = (0.06, 0.43), $r^2 = 0.07$]. Examining transitions between post-originality shows that these effects stem from slightly different posting patterns. Users with higher CES-D scores tend to follow non-original content with non-original content [N↔N, *post-level with-silence*, $r = 0.26$, $p < 0.05$, 95% CI = (0.07, 0.43), $r^2 = 0.07$] or to alternate between original and non-original content [N↔O *post-level with-silence*, $r = 0.27$, $p < 0.05$, 95% CI = (0.08, 0.44), $r^2 = 0.07$]. Users with higher neuroticism scores tend to post-sequences of non-original content [N↔N, post-level with-silence, $r = 0.25$, $p < 0.05$, 95% CI = (0.06, 0.43), $r^2 = 0.05$] and are less likely to post-original content before or after a period of silence [O↔S, *post-level with-silence*, $r = 0.28$, $p < 0.05$, 95% CI = (0.09, 0.45), $r^2 = 0.08$].

Since neuroticism is closely linked to depression symptoms, we also computed a partial correlation between content originality and CES-D while controlling for neuroticism. The resulting correlation was no longer significant ($r = 0.14$, $p = 0.22$, $r^2 = 0.02$). Therefore, the association between content originality and depression symptoms might be moderated by neuroticism.

5. DISCUSSION

5.1. Main Findings

Many studies have found associations between the frequency of affective words used in social media text and personality. However, existing studies often saw affect as static and only focused on the strength of bipolar valence (positive/negative). Instead, our work focuses on affect patterns. We encode posting behavior, transitions between affect states, and content originality. From a practical point of view, our technique can supplement experience sampling techniques (Myin-Germeys

TABLE 3 | Correlations between personality, SWL, and CES-D scores and affect transitions. Number of participants $N = 70$.

	S↔S	−↔−	+↔+	±↔± ±↔±	0↔0	+↔−	±↔+	±↔−	±↔0	0↔+	0↔−	S↔+	S↔−	±↔S	S↔0
Post-level representation (post-plus silence)															
N_{Occ}	1238	346	542	29	230	599	143	134	100	424	414	641	384	137	211
ope	0.09	−0.17	−0.17	−0.16	−0.05	−0.14	−0.07	−0.08	0.11	0.01	0.03	0.17	0.00	0.13	0.03
con	−0.06	0.01	0.09	−0.09	−0.15	0.11	0.00	−0.01	−0.14	−0.07	−0.08	0.16	0.00	0.15	−0.15
ext	0.04	−0.12	0.16	−0.10	−0.19	−0.06	−0.03	−0.12	−0.09	−0.09	−0.17	**0.29****	−0.04	0.00	−0.18
agr	0.14	**−0.37*****	0.03	0.02	−0.15	**−0.22***	0.08	0.04	0.04	−0.04	**−0.23***	**0.23***	−0.04	**0.29****	−0.13
neu	−0.07	0.19	0.18	0.18	−0.03	**0.23***	0.11	0.04	0.02	0.05	−0.05	**−0.22***	−0.03	**−0.23***	−0.13
swl	0.04	−0.10	−0.13	−0.10	0.06	−0.03	0.02	−0.05	−0.04	0.02	−0.08	0.02	0.16	−0.02	0.18
CESD	−0.04	0.19	0.08	0.09	0.00	0.04	0.15	0.07	0.03	−0.06	0.11	−0.20·	0.00	−0.11	−0.03
Post-level representation (post-only), $N = 70$															
N_{Occ}		396	694	34	313	728	188	166	142	547	502				
ope		−0.16	−0.05	−0.06	−0.02	−0.05	0.06	−0.01	0.14	0.09	0.13				
con		−0.07	0.18	−0.07	**−0.23***	0.08	0.14	0.10	−0.11	−0.13	−0.12				
ext		−0.04	**0.33*****	0.04	**−0.24***	0.05	0.08	−0.10	−0.15	−0.16	−0.20·				
agr		**−0.28****	0.18	0.00	−0.16	−0.10	**0.26***	**0.28****	0.13	0.03	**−0.26***				
neu		0.14	0.00	0.11	−0.02	0.16	−0.14	−0.09	−0.08	0.01	−0.12				
swl		0.00	−0.12	−0.11	0.11	0.02	0.09	0.09	−0.02	0.08	−0.04				
CESD		0.14	−0.04	0.03	0.04	−0.03	−0.06	−0.11	0.04	−0.11	0.13				
Day-level representation, $N = 70$															
N_{Occ}	228	281	271	267	304	287	303	296	298	261	311	242	259	261	261
ope	0.12	−0.17	−0.11	−0.05	−0.02	−0.08	0.00	−0.14	−0.07	−0.01	−0.02	0.12	−0.02	0.19	0.13
con	−0.06	−0.03	**0.25***	0.05	−0.01	0.03	−0.03	−0.04	−0.16	−0.19	−0.12	0.08	0.10	0.06	−0.07
ext	0.06	−0.11	**0.30*****	−0.03	−0.14	0.04	0.14	−0.13	0.01	**−0.28****	**−0.28****	**0.24***	−0.08	0.02	−0.17
agr	0.11	−0.22	0.15	−0.05	0.08	−0.12	0.16	−0.06	0.11	−0.08	−0.17	0.15	−0.07	**0.28****	−0.09
neu	−0.08	0.16	0.00	0.19	−0.17	**0.21***	0.09	0.11	−0.01	0.12	0.08	−0.14	−0.12	**−0.26***	−0.03
swl	0.02	−0.08	−0.01	−0.08	**0.25***	−0.03	−0.06	−0.10	0.03	−0.06	−0.04	−0.02	0.12	0.06	0.08
CESD	−0.03	0.11	−0.10	0.08	−0.18	0.02	0.10	0.08	0.08	−0.01	0.21·	−0.18	0.03	−0.16	0.05

*Pearson correlation P-value (permutation testing): · < 0.1, * < 0.05, ** < 0.01, *** < 0.001, bidirectional transition types: ↔; −, negative valence; +, positive valence; ±, mixed valence; 0, neutral; S, silence day; N_{Occ}, number of occurrences of each transition type; ope, openness; con, conscientiousness; ext, extraversion; agr, agreeableness; neu, neuroticism; swl, satisfaction with life scale; CESD, Center for Epidemiologic Studies Depression Scale. Bold: $p < 0.05$.*

et al., 2018) to help clinicians and patients develop a more comprehensive view of a person's affect patterns, arrive at a better-substantiated diagnosis, and make improved treatment decisions. However, this depends on whether the patient is willing to share information from their social media feed with their therapist. Overall, the correlations seen between affect transitions and personality traits are in line with the consensus in the early literature (Gross et al., 1998). Extroverts tend to produce sequences of positive posts. This behavior fits well with the positive emotional core in extroverts stipulated in (Watson and Clark, 1997). Participants with higher agreeableness are less likely to post-sequences of negative posts. This could be due to their ability to regulate negative affect (Meier et al., 2006; Haas et al., 2007).

Although the psychology literature suggests a strong association between negative mood states and neuroticism (Rusting and Larsen, 1995), we did not find this in our data. Our results are in line with previous studies of verbal cues to personality traits in social media (Yarkoni, 2010; Golbeck et al., 2011; Schwartz et al., 2013; Park et al., 2015). Golbeck et al. (2011) found that social media users who were more likely to talk about anxiety were on the higher end of the neuroticism scale. We speculate that self-presentation bias may influence how social media users regulate their expression of negative emotions in their public posts. The only relevant association we found was that social media users on the high end of neuroticism are more likely to switch between posting positive and negative affective content. This finding aligns well with the

TABLE 4 | Summary of the significant correlations between transition states and the five personality traits ($p < 0.01$).

	Transitions	Post-level (with-silence)	Post-level (without-silence)	Day-level
Extraversion	S ↔ +	↑	–	–
	0 ↔ +	–	–	↓
	0 ↔ –	–	–	↓
	+ ↔ +	–	↑	↑
Agreeablness	– ↔ –	↓	↓	–
	± ↔ S	↑	–	↑
	± ↔ –	–	↑	–

↓ Indicates a significant negative correlation at $p<0.01$ or better, ↑ indicates a significant positive correlation at $p<0.01$ or better. — Indicates that the correlation is not significant at this level. Bidirectional transition types: ↔; –, negative valence; +, positive valence; ±, mixed valence; 0, neutral; S, silence day.

fact that high neuroticism is associated with high emotional instability (Costa and McCrae, 1992).

The link between posting non-original content and elevated depression symptoms appears to be moderated by neuroticism. This suggests that high levels of neuroticism predispose users both to depressive symptoms and to an indirect disclosure of emotions through quotes and lyrics.

In our sample, the prevalence of depressive symptoms is higher than would be expected in the general population. In the original CES-D paper, Radloff (1977) proposed three levels of depression severity: low (0–15), mild-to-moderate (16–22), and high (23–60). They found that only 21% of the general population scored above the low symptom level. In contrast, in our sample, nearly half of the participants exhibit a high level of symptoms (>22). Within the context of social media studies of depression, however, our data set is not exceptional. For many studies in the area, high symptom individuals account for nearly half of the data set (De Choudhury et al., 2013; Tsugawa et al., 2015; Nadeem, 2016; Reece et al., 2017; Orabi et al., 2018).

Our results support the claim that affect expressed in social media data text is associated with social media users' affect patterns in real life. However, the data set used in this study is from the early 2010's and only covers the well-established social media platform Facebook. The associations found in this study are likely to be slightly different from those found in another social networks (e.g., Instagram) or in a new data set collected 10 years later.

5.2. Limitations

Due to the restrictions imposed by the need for sufficient Facebook updates to allow analysis, our final sample is relatively small. Given the size of the significant effects we found in the data, power calculations indicate that a well-powered study should include data from around 200 users (Schönbrodt and Perugini, 2013). It also skews heavily toward younger female Caucasians with relatively low satisfaction with life and strong depression symptoms. It is possible that other groups of users (e.g., non-Caucasians, males) are less likely to disclose personal information about mood and emotions on their public Facebook pages (Dosono et al., 2017; McDonald et al., 2019).

6. CONCLUSION AND FUTURE WORK

In this pilot study, we demonstrated the benefits of detailed representations of social media affect for unpacking the relationship between personality, mental well-being, and the content posted on social media. Importantly, our representations include non-binary affect categories (positive, negative, mixed, neutral), and take into account content originality. As a consequence, we were able to obtain a more detailed picture of the link between patterns of affect and depressive symptoms.

In future work, we plan to enrich our data set with more in-depth analyses of original vs. non-original content, extend coverage by including a larger sample of the myPersonality data set, and construct statistical models that allow us to observe long-term trends in posting patterns. Future studies should also examine the extent to which affect expressed in non-original content is aligned with the users' affect when they post the material.

DATA AVAILABILITY STATEMENT

The datasets generated for this study will not be made publicly available because the myPersonality database is closed for further research. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Self-Certification according to the procedure of the School of Informatics, University of Edinburgh. The patients/participants provided their written informed consent to participate in this study. The secondary analysis of this data set was reviewed and approved by the Ethics Committee of the School of Informatics, University of Edinburgh, Reference Number 72771.

AUTHOR CONTRIBUTIONS

LC: study design, statistical analysis, analysis of results, and drafting of paper. WM: principal supervisor of LC. MW: second supervisor of LC. WM and MK contributed to paper writing, advised on study design, statistical analysis, and analysis of results.

ACKNOWLEDGMENTS

We thank Michael Kosinski and David Stilwell for permission to use myPersonality, and our six undergraduate Research Assistants from the Psychology Department of the University of Edinburgh for their hard annotation work. The work of WM and MW on this paper was partly funded by The Alan Turing Institute (EPSRC, EP/N510129/1).

REFERENCES

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011). "Sentiment analysis of twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)* (Portland, OR), 30–38.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., and Stillwell, D. (2012). "Personality and patterns of facebook usage," in *Proceedings of the 4th Annual ACM Web Science Conference* (New York, NY: ACM), 24–32. doi: 10.1145/2380718.2380722

Balahur, A., Steinberger, R., Goot, E. v. d., Pouliquen, B., and Kabadjov, M. (2009). "Opinion mining on newspaper quotations," in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 3 (Milano: IEEE Computer Society), 523–526. doi: 10.1109/WI-IAT.2009.340

Bazarova, N. N., Choi, Y. H., Schwanda Sosik, V., Cosley, D., and Whitlock, J. (2015). "Social sharing of emotions on facebook: channel differences, satisfaction, and replies," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC: ACM), 154–164. doi: 10.1145/2675133.2675297

Bylsma, L. M., Morris, B. H., and Rottenberg, J. (2008). A meta-analysis of emotional reactivity in major depressive disorder. *Clin. Psychol. Rev*. 28, 676–691. doi: 10.1016/j.cpr.2007.10.001

Bylsma, L. M., Taylor-Clift, A., and Rottenberg, J. (2011). Emotional reactivity to daily events in major and minor depression. *J. Abnorm. Psychol*. 120:155. doi: 10.1037/a0021662

Carlo, G., Mestre, M. V., McGinley, M. M., Samper, P., Tur, A., and Sandman, D. (2012). The interplay of emotional instability, empathy, and coping on prosocial and aggressive behaviors. *Pers. Individ. Differ*. 53, 675–680. doi: 10.1016/j.paid.2012.05.022

Chen, L., Cheng, C. H. K., and Gong, T. (2020). Inspecting vulnerability to depression from social media affect. *Front. Psychiatry* 11:54. doi: 10.3389/fpsyt.2020.00054

Chen, L., Gong, T., Kosinski, M., Stillwell, D., and Davidson, R. L. (2017). Building a profile of subjective well-being for social media users. *PLoS ONE* 12:e0187278. doi: 10.1371/journal.pone.0187278

Coppersmith, G., Dredze, M., and Harman, C. (2014). "Quantifying mental health signals in twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Denver, CO), 51–60. doi: 10.3115/v1/W14-3207

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., and Mitchell, M. (2015). "Clpsych 2015 shared task: depression and ptsd on twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Denver, CO), 31–39. doi: 10.3115/v1/W15-1204

Costa, P. T., and McCrae, R. R. (1992). *Neo Pi-R*. Odessa, FL: Psychological Assessment Resources.

De Choudhury, M., and De, S. (2014). "Mental health discourse on reddit: Self-disclosure, social support, and anonymity," in *Eighth International AAAI Conference on Weblogs and Social Media* (Michigan), 21–30.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). "Predicting depression via social media," in *Seventh International AAAI Conference on Weblogs and Social Media* (Cambridge, MA), 170–185.

Diener, E., and Emmons, R. A. (1984). The independence of positive and negative affect. *J. Pers. Soc. Psychol*. 47:1105. doi: 10.1037/0022-3514.47.5.1105

Diener, E., Suh, E. M., Lucas, R. E., and Smith, H. L. (1999). Subjective well-being: three decades of progress. *Psychol. Bull*. 125:276. doi: 10.1037/0033-2909.125.2.276

Diener, E. D., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess*. 49, 71–75. doi: 10.1207/s15327752jpa4901_13

Digman, J. M. (1990). Personality structure: emergence of the five-factor model. *Annu. Rev. Psychol*. 41, 417–440. doi: 10.1146/annurev.ps.41.020190.002221

Dosono, B., Rashidi, Y., Akter, T., Semaan, B., and Kapadia, A. (2017). Challenges in transitioning from civil to military culture: hyper-selective disclosure through ICTs. *Proc. ACM Hum. Comput. Interact*. 1, 41:1–41:23. doi: 10.1145/3134676

Farnadi, G., Zoghbi, S., Moens, M.-F., and De Cock, M. (2013). "Recognising personality traits using facebook status updates," in *Seventh International AAAI Conference on Weblogs and Social Media* (Cambridge, MA), 154–164.

Fredrickson, B. L. (1998). What good are positive emotions? *Rev. Gen. Psychol*. 2, 300–319. doi: 10.1037/1089-2680.2.3.300

Frijda, N. H. (1993). Moods, emotion episodes, and emotions. *Handb. Emot*. 12:155.

Fujita, F., Diener, E., and Sandvik, E. (1991). Gender differences in negative affect and well-being: the case for emotional intensity. *J. Pers. Soc. Psychol*. 61:427. doi: 10.1037/0022-3514.61.3.427

Gamer, M., Lemon, J., Fellows, I., and Singh, P. (2019). *irr: Various Coefficients of Inter-Rater Reliability and Agreement*. R package version 0.84.81. Available online at: https://cran.r-project.org/web/packages/irr/irr.pdf

Garren, S. T. (2017). Permutation Tests for Nonparametric Statistics Using R. *Asian J. Math*. 5, 1–8.

Golbeck, J., Robles, C., Edmondson, M., and Turner, K. (2011). "Predicting personality from twitter," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (IEEE), 149–156. doi: 10.1109/PASSAT/SocialCom.2011.33

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *J. Res. Pers*. 40, 84–96. doi: 10.1016/j.jrp.2005.08.007

Gross, J. J., Sutton, S. K., and Ketelaar, T. (1998). Relations between affect and personality: support for the affect-level and affective-reactivity views. *Pers. Soc. Psychol. Bull*. 24, 279–288. doi: 10.1177/0146167298243005

Haas, B. W., Omura, K., Constable, R. T., and Canli, T. (2007). Is automatic emotion regulation associated with agreeableness? A perspective using a social neuroscience approach. *Psychol. Sci*. 18, 130–132. doi: 10.1111/j.1467-9280.2007.01861.x

Headey, B., Kelley, J., and Wearing, A. (1993). Dimensions of mental health: life satisfaction, positive affect, anxiety and depression. *Soc. Indicat. Res*. 29, 63–82. doi: 10.1007/BF01136197

Higgins, J. (2003). Introduction to Modern Non-Parametric Statistics. *The American Statistician*. Pacific Grove, CA, 61:184. doi: 10.1198/tas.2007.s81

Houben, M., Van Den Noortgate, W., and Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: a meta-analysis. *Psychol. Bull*. 141:901. doi: 10.1037/a0038822

Juslin, P. N., and Laukka, P. (2004). Expression, perception, and induction of musical emotions: a review and a questionnaire study of everyday listening. *J. New Music Res*. 33, 217–238. doi: 10.1080/0929821042000317813

Lin, J., Mao, W., and Zeng, D. D. (2017). Personality-based refinement for sentiment classification in microblog. *Knowl. Based Syst*. 132, 204–214. doi: 10.1016/j.knosys.2017.06.031

McCrae, R. R., and Allik, J. (Eds.). (2002). *The Five-Factor Model of Personality Across Cultures*. New York, NY: Springer Science and Business Media.

McCrae, R. R., and John, O. P. (1992). An introduction to the five–factor model and its applications. *J. Pers*. 60, 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

McDonald, J., Harris, K. L., and Ramirez, J. (2019). Revealing and concealing difference: a critical approach to disclosure and an intersectional theory of "closeting". *Commun. Theory* 30, 84–104. doi: 10.1093/ct/qtz017

Meier, B. P., Robinson, M. D., and Wilkowski, B. M. (2006). Turning the other cheek: agreeableness and the regulation of aggression-related primes. *Psychol. Sci*. 17, 136–142. doi: 10.1111/j.1467-9280.2006.01676.x

Mervielde, I., Buyst, V., and Fruyt, F. D. (1995). The validity of the big-five as a model for teachers' ratings of individual differences among children aged 4–12 years. *Pers. Individ. Differ*. 18, 525–534. doi: 10.1016/0191-8869(94)00175-R

Mohammad, S. (2016). "A practical guide to sentiment annotation: challenges and solutions," in *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (San Diego, CA), 174–179. doi: 10.18653/v1/W16-0429

Moilanen, K., and Pulman, S. (2007). "Sentiment composition," in *Proceedings of the Recent Advances in Natural Language Processing International Conference* (Borovets), 378–382.

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., et al. (2018). Experience sampling methodology in mental health research:

new insights and technical developments. *World Psychiatry* 17, 123–132. doi: 10.1002/wps.20513

Nadeem, M. (2016). Identifying depression on twitter. *arXiv* 1607.07384.

Orabi, A. H., Buddhitha, P., Orabi, M. H., and Inkpen, D. (2018). "Deep learning for depression detection of twitter users," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (New Orleans, LA), 88–97.

Orme, J. G., Reis, J., and Herz, E. J. (1986). Factorial and discriminant validity of the center for epidemiological studies depression (ces-d) scale. *J. Clin. Psychol.* 42, 28–33. doi: 10.1002/1097-4679(198601)42:1<28::AID-JCLP2270420104>3.0.CO;2-T

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., et al. (2015). Automatic personality assessment through social media language. *J. Pers. Soc. Psychol.* 108:934. doi: 10.1037/pspp0000020

Park, M., Cha, C., and Cha, M. (2012). "Depressive moods of users portrayed in twitter," in *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)*, Vol. 2012 (New York, NY: ACM), 1–8.

Pavot, W., and Diener, E. (2009). Review of the satisfaction with life scale. *Psychol. Assess.* 5:164. doi: 10.1037/1040-3590.5.2.164

Pennebaker, J. W., and King, L. A. (1999). Linguistic styles: Language use as an individual difference. *J. Pers. Soc. Psychol.* 77:1296. doi: 10.1037/0022-3514.77.6.1296

Pishva, N., Ghalehban, M., Moradi, A., and Hoseini, L. (2011). Personality and happiness. *Proc. Soc. Behav. Sci.* 30, 429–432. doi: 10.1016/j.sbspro.2011.10.084

Radloff, L. S. (1977). The CES-D scale: a self-report depression scale for research in the general population. *Appl. Psychol. Meas.* 1, 385–401. doi: 10.1177/014662167700100306

Reece, A. G., Reagan, A. J., Lix, K. L., Dodds, P. S., Danforth, C. M., and Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data. *Sci. Rep.* 7:13006. doi: 10.1038/s41598-017-12961-9

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., and Boyd-Graber, J. (2015). "Beyond LDA: exploring supervised topic modeling for depression-related language in twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Denver, Co), 99–107. doi: 10.3115/v1/W15-1212

Roberts, R. E. (1980). Reliability of the ces-d scale in different ethnic contexts. *Psychiatry Res.* 2, 125–134. doi: 10.1016/0165-1781(80)90069-4

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). "Semeval-2015 task 10: sentiment analysis in twitter," in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)* (Denver, CO), 451–463. doi: 10.18653/v1/S15-2078

Rothbart, M. K., Ahadi, S. A., and Evans, D. E. (2000). Temperament and personality: origins and outcomes. *J. Pers. Soc. Psychol.* 78:122. doi: 10.1037/0022-3514.78.1.122

Rottenberg, J. (2005). Mood and emotion in major depression. *Curr. Direct. Psychol. Sci.* 14, 167–170. doi: 10.1111/j.0963-7214.2005.00354.x

Rottenberg, J., and Gross, J. J. (2003). When emotion goes wrong: realizing the promise of affective science. *Clin. Psychol. Sci. Pract.* 10, 227–232. doi: 10.1093/clipsy.bpg012

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39:1161. doi: 10.1037/h0077714

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychol. Rev.* 110:145. doi: 10.1037/0033-295X.110.1.145

Rusting, C. L. (1998). Personality, mood, and cognitive processing of emotional information: three conceptual frameworks. *Psychol. Bull.* 124:165. doi: 10.1037/0033-2909.124.2.165

Rusting, C. L., and Larsen, R. J. (1995). Moods as sources of stimulation: relationships between personality and desired mood states. *Pers. Individ. Differ.* 18, 321–329. doi: 10.1016/0191-8869(94)00157-N

Ryan, R. M., and Deci, E. L. (2001). On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annu. Rev. Psychol.* 52, 141–166. doi: 10.1146/annurev.psych.52.1.141

Scherer, K. R. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *J. New Music Res.* 33, 239–251. doi: 10.1080/0929821042000317822

Scherer, K. R., and Zentner, M. R. (2001). Emotional effects of music: production rules. *Music Emot. Theory Res.* 361:392.

Schimmack, U., Oishi, S., and Diener, E. (2002). Cultural influences on the relation between pleasant emotions and unpleasant emotions: Asian dialectic philosophies or individualism-collectivism? *Cogn. Emot.* 16, 705–719. doi: 10.1080/02699930143000590

Schönbrodt, F. D., and Perugini, M. (2013). At what sample size do correlations stabilize? *J. Res. Pers.* 47, 609–612. doi: 10.1016/j.jrp.2013.05.009

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS ONE* 8:e0073791. doi: 10.1371/journal.pone.0073791

Schwartz, S. H. (1992). "Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries," in *Advances in Experimental Social Psychology*, Vol. 25 (Cambridge, MA: Elsevier), 1–65. doi: 10.1016/S0065-2601(08)60281-6

Sheppes, G., Suri, G., and Gross, J. J. (2015). Emotion regulation and psychopathology. *Annu. Rev. Clin. Psychol.* 11, 379–405. doi: 10.1146/annurev-clinpsy-032814-112739

Silvera, D. H., Lavack, A. M., and Kropp, F. (2008). Impulse buying: the role of affect, social influence, and subjective wellbeing. *J. Consum. Market.* 25, 23–33. doi: 10.1108/07363760810845381

Singh, K., and Jha, S. D. (2008). Positive and negative affect, and grit as predictors of happiness and life satisfaction. *J. Indian Acad. Appl. Psychol.* 34, 40–45.

Teufel, S. (1999). *Argumentative zoning: information extraction from scientific articles* (Ph.D. thesis), Centre for Cognitive Science, University of Edinburgh, Edinburgh, United Kingdom.

Thompson, R. J., Mata, J., Jaeggi, S. M., Buschkuehl, M., Jonides, J., and Gotlib, I. H. (2012). The everyday emotional experience of adults with major depressive disorder: examining emotional instability, inertia, and reactivity. *J. Abnorm. Psychol.* 121:819. doi: 10.1037/a0027978

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., and Ohsaki, H. (2015). "Recognizing depression from twitter activity," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM), 3187–3196. doi: 10.1145/2702123.2702280

Watson, D. (2000). *Mood and Temperament*. New York, NY: Guilford Press.

Watson, D., and Clark, L. A. (1997). "Extraversion and its positive emotional core," in *Handbook of Personality Psychology* (Amsterdam: Elsevier), 767–793. doi: 10.1016/B978-012134645-4/50030-5

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

Yarkoni, T. (2010). Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers. *J. Res. Pers.* 44, 363–373. doi: 10.1016/j.jrp.2010.04.001

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1036–1040. doi: 10.1073/pnas.1418680112

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 June 2020
doi: 10.3389/fpsyg.2020.01321

The Priest, the Sex Worker, and the CEO: Measuring Motivation by Job Type

Jan Ketil Arnulf[1], Kim Nimon[2], Kai Rune Larsen[3], Christiane V. Hovland[1] and Merethe Arnesen[1]*

[1] Department of Leadership and Organizational Behaviour, BI Norwegian Business School, Oslo, Norway, [2] Department of Human Resource Development, The University of Texas at Tyler, Tyler, TX, United States, [3] Leeds Business School, University of Colorado at Boulder, Boulder, CO, United States

This study uses latent semantic analysis (LSA) to explore how prevalent measures of motivation are interpreted across very diverse job types. Building on the Semantic Theory of Survey Response (STSR), we calculate "semantic compliance" as the degree to which an individual's responses follow a semantically predictable pattern. This allows us to examine how context, in the form of job type, influences respondent interpretations of items. In total, 399 respondents from 18 widely different job types (from CEOs through lawyers, priests and artists to sex workers and professional soldiers) self-rated their work motivation on eight commonly applied scales from research on motivation. A second sample served as an external evaluation panel (n = 30) and rated the 18 job types across eight job characteristics. Independent measures of the job types' salary levels were obtained from national statistics. The findings indicate that while job type predicts motivational score levels significantly, semantic compliance as moderated by job type job also predicts motivational score levels usually at a lesser but significant magnitude. Combined, semantic compliance and job type explained up to 41% of the differences in motional score levels. The variation in semantic compliance was also significantly related to job characteristics as rated by an external panel, and to national income levels. Our findings indicate that people in different contexts interpret items differently to a degree that substantially affects their score levels. We discuss how future measurements of motivation may improve by taking semantic compliance and the STSR perspective into consideration.

Keywords: motivation, semantic theory of survey response, Likert scale analysis, job types, job design theory, self-determination theory, latent semantic analysis

OPEN ACCESS

Edited by:
Roberto Codella, University of Milan, Italy

Reviewed by:
Andrea Chirico, Sapienza University of Rome, Italy
Francesco Giancamilli, Sapienza University of Rome, Italy

****Correspondence:***
Jan Ketil Arnulf
jan.k.arnulf@bi.no

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

Received: *06 February 2020*
Accepted: *19 May 2020*
Published: *19 June 2020*

Citation:
Arnulf JK, Nimon K, Larsen KR, Hovland CV and Arnesen M (2020) The Priest, the Sex Worker, and the CEO: Measuring Motivation by Job Type. Front. Psychol. 11:1321. doi: 10.3389/fpsyg.2020.01321

INTRODUCTION

> "Most social acts have to be understood in their setting, and lose meaning if isolated. No error in thinking about social facts is more serious than the failure to see their place and function" Asch (1987).

Asch's (1987, p. 61, orig. 1952) warning is as relevant today as half a century ago. The numbers emerging from Likert-scale data are what social anthropologist Geertz (1973) called "thin data" because they reduce a complex experience to seemingly uniform rows of numbers. The meaning of these numbers is still debated in the methodological literature (Drasgow et al., 2015;

Kjell et al., 2019). From a linguistic point of view, it is unlikely that short sentences of the type normally used in Likert-scale items will mean the same to all people regardless of the context of the respondents (Kay, 1996; Borsboom, 2008; Maul, 2017). To the extent that people interpret items differently according to their own situations, the item texts function like a story, where the items combine in different ways to describe different contexts.

This study explores how people responding to the same items about motivation seem to interpret these in different ways dependent on their professional contexts, a phenomenon not accounted for in most theories, and not part of standard psychometrics. Our aim is to show how item interpretation may be almost be as deep a characteristic of different groups as the score levels themselves.

Recent developments in quantitative text analysis suggest that quantitative responses to survey items may be heavily influenced by semantics (e.g., Nimon et al., 2016; Rosenbusch et al., 2019; Arnulf and Larsen, 2020). The Semantic Theory of Survey Response (STSR) claims that the most obvious reason for covariation between items is that they are semantically related (Arnulf et al., 2018d). Empirical testing of STSR has revealed that the correlation matrix of survey data can be strongly determined by their semantic properties, but not always, and not necessarily to the same extent across all groups (e.g., Arnulf et al., 2014; Nimon et al., 2016). In the present study, we examine whether interpretation of the same item sets differs systematically across contextually consistent respondent sets. For example, will items on the motivational effects of payments mean the same regardless of the expected income of people?

A study by Drasgow et al. (2015) expressed doubts about interpreting Likert-scale measurements as "dominant measures" where all traits are uniformly scalable from low to high. Instead, they suggested that respondents display preferred values, choosing alternatives that more accurately describe their viewpoints but not necessarily on a more-or-less continuum. A similar argument has been raised in a study that used semantic algorithms to rate free-text responses in a personality survey (Kjell et al., 2019).

The purpose of this study is therefore to explore the degree to which subjects from different professional contexts respond to motivational items in ways that cohere with or deviate from what is semantically expected. The contributions of this study are to: (a) strengthen STSR by establishing a technique for assessing the mutual impact of score levels and semantic characteristics of items in differentiating between groups of respondents, (b) contribute a general understanding of the psychology involved in item responses for different occupational groups, and (c) advance ways to use semantic algorithms as a methodological tool in social sciences including and not limited to organizational behavior and social psychology.

THEORY

In his original description of the scales that now carry his name, Likert (1932, p. 7, italics in orig.) wrote: "...it is strictly true that the number of attitudes which any given person possesses is almost infinite. This result is statistically as well as psychologically absurd. Exactly the same absurdity and the same obstacle to research is offered by those definitions of attitude which conceive them merely as *verbal expressions*..."

Now, almost 100 years later, working with the verbal expressions is no longer an absurdity, neither statistically nor psychologically. Language algorithms have opened a way to work precisely with the self-descriptive statements that Likert (1932) and his contemporaries could not address (Nimon et al., 2016; Arnulf et al., 2018a; Kjell et al., 2019). Our basic assumption in this study builds on the linguistic fact that all worded statements mean different things to different subjects dependent on their context (Kay, 1996; Sidnell and Enfield, 2012). An example of this has previously been described by Putnick and Bornstein (2016) who noted that symptoms of depression such as crying are different between men and women, influencing the score levels on questions with such content. Similarly, our focus here is on how questions about motivation may take on different meanings in different job types, affecting score levels.

In what follows, we suggest that the common approach to treating survey responses as measures builds on an incomplete understanding of the meaning of the numbers with respect to their semantic dependencies. We then argue that semantic analysis is a viable approach to a different appreciation of survey items that may possibly alleviate some of the previously described problems. The arguments will be tested in an empirical analysis of a dataset containing self-rated motivation across very different professional contexts to support our claims. To finally ascertain that the semantic influence is not methodological artifact, we will validate the semantic data with two other independent data sources, an independent rating panel and national income statistics.

Likert Scale Measures of Contextual Motivation

Likert's (1932) argument for trusting the numbers from his scales was that working with verbal expressions would be methodologically impossible. Hence, the simplification of attitudes imposed on responses by using numerical scales was the only workable solution for empirical research. To this day, history has judged Likert right and the use of his scales is one of the most commonly used measurement instruments in social science research and enables a range of practical applications (Likert, 1932; Podsakoff et al., 2003; Sirota et al., 2005; Cascio, 2012; Yukl, 2012a; Lamiell, 2013).

Despite or perhaps even because of its intuitive simplicity, however, other researchers have been critical of some of the uses of Likert scales since the days of its conception (e.g., Andrich, 1996; Drasgow et al., 2015). One problematic aspect of Likert scales is that while the response categories are usually framed as texts, they are transformed into numbers used for calculations. These numbers are in turn translated back to texts as inferences about the measured attitudes (Kjell et al., 2019). An item with the text "I will look for a new job in the next year" may be scored as "Definitely not – probably not – maybe – probably – definitely yes." The choice of the option "I will definitely look

for a new job in the next year" would be assigned a numerical value (e.g., 5) as a measure used for calculations. Following a commonly used convention, "measurement" can be defined as the "process of assigning numbers to represent qualities" (Campbell, 1920, p. 267). However, it is not entirely obvious what numbers from Likert scales measure (Smedslund, 1988; Elster, 2011; Mari et al., 2017; Slaney, 2017). While measurement is a complex concept that can be defined in numerous ways, some conventionality in the definition of measurement seems unavoidable (Mari et al., 2017, pp. 117–121). The conventionality or common sense element seems to require that a "measure" should retain its meaning across contexts in order to be a valid measurement. We expect measurement units of walls and floors to be consistent independently of the size of buildings and expect temperature assessments to allow comparisons of polar with tropical environments. Such invariance does not necessarily apply to numbers from Likert scale items. "A warm day" refers to very different measured temperatures in Texas and Norway, and used as a survey item, the distinction between contexts may blur. The same problem could possibly arise with measurements in social science. Will the same statement about motivation imply the same attitudinal measure across contexts? Or will "satisfaction with pay" mean different things dependent on the difference in payment levels between job types?

The STSR offers a framework to test these questions empirically. The theory posits that for two different items to be scored independently, they also need to be semantically independent. If two items are semantically intertwined, the answer to the second will somehow depend on the first – unless the respondents make different interpretations of the items (Schwarz, 1999). It is this difference that we can try to assess using the semantic techniques that we will explain below.

Motivation is a latent variable (Borsboom, 2008). Assessments of *motivational strength* are therefore not directly accessible, even to the individual in question (McClelland et al., 1989; Parks-Stamm et al., 2010). Thus, self-rated motivation is likely to be influenced by a number of factors. However, a large number of studies on motivation in the workplace have relied on Likert-scale items to model motivational effects. Among these studies, two theories stand out as particularly relevant to our aim: The Job Characteristics Model (JCM) originally proposed by Hackman and Oldham (1976) proposed that different job contexts – their characteristics – would have systematically different impacts on employee motivation. A later development, the Self-Determination Theory (SDT) built on this and outlined how contextual variables could translate into types of motivation that enhance or impair performance (Deci et al., 1989, 2017). Building on these traditions, Barrick et al. (2013) outlined how individual characteristics interact with situational variables in a sense-making process to create different types of job motivation through experiencing work as meaningful. We can thus build a framework of theory and existing research to assess the impact of semantics on survey responses in work motivation:

Our first interest concerns work contexts as we assume that these will impact motivational levels *as well as* the interpretations of items. The Job Characteristics Model (JCM) (Hackman and Oldham, 1975, 1976) has been in prevalent use for work design over two decades (Kanfer et al., 2017, p. 342). Precisely because JCM focuses on *job characteristics,* the model should help us identify aspects of jobs that are inter-subjectively valid and not indicative of individual differences between employees. In fact, the origins of JCM was an explicit intention to identify situational variables such that one may measure the impact of job design on motivation (Hackman and Oldham, 1976, p. 252).

According to JCM, five *core* job dimensions will affect motivation: (a) skill variety, (b) task identity, (c) task significance, (d) autonomy, and (e) feedback (Hackman and Oldham, 1975, 1976). We therefore assume that these dimensions will be important descriptors of jobs where subjects may vary in types and levels of motivation as well as in their interpretation of items. Expanding on these, later research has identified enriched social roles, influence and status as belonging to taxonomies of job situations (Oldham and Hackman, 2010; Barrick et al., 2013).

The JCM theory presumes that job characteristics will interact with different needs in the different employees to induce levels of motivation (Hackman and Oldham, 1976). This subjective interpretive process has been elaborated in more detail by Self-Determination Theory (SDT) (Deci et al., 1989, 2017; Ryan and Deci, 2000a,b), and has served as framework for research on motivation and work outcomes using self-perception with Likert-type rating scales (Grant, 2008; Dysvik et al., 2010; Fang and Gerhart, 2012; Rockmann and Ballinger, 2017).

According to SDT, conditions that activate motivation can be distinguished on a continuum from autonomous to controlled, where controlled types of motivation are less favorable: "external regulation can powerfully motivate specific behaviors, but it often comes with collateral damage in the form of long-term decrements in autonomous motivation and well-being, sometimes with organizational spillover effects" (Deci et al., 2017, p. 21). Instead, autonomous motivation – where intrinsic motivation (IM) or pleasure in the activity for its own sake is one type, tends to have better outcomes: "Employees can be intrinsically motivated for at least parts of their jobs, if not for all aspects of them, and when intrinsically motivated the individuals tend to display high-quality performance and wellness" (Deci et al., 2017, p. 21).

As can be seen from the explanations above, SDT does not assume an automatic relationship between situational context and type of motivation. Rather, the sub-optimal effect of extrinsic motivation is linked to a perception of being controlled. Also, IM is not always assumed to induce superior performance to extrinsic motivation. Still, the aim of the theory is to guide managerial practices that facilitate intrinsic types of motivation, because these are generally seen to produce better outcomes. The relationship to situations is clearly outlined in a recent summary of research in the field (Deci et al., 2017, p. 20): "Some have careers that are relatively interesting and valued by others. Their work conditions are supportive, and they perceive their pay to be equitable. Others, however, have jobs that are demanding and demeaning. Their work conditions are uncomfortable, and their pay is not adequate for supporting a family. They are likely to look forward to days away from work to feel alive and well." The cited summary reviews a number of studies that show how extrinsic rewards may reduce performance through experience of being

controlled, and how IM generally leads to better performance in terms of effort, quality, and subjective wellness.

The final point to be elaborated is the interpretive process that translates the job characteristics into the experienced motivational states. Outlining a "theory of purposeful behavior," Barrick et al. (2013, p. 149) claimed that individuals take an agentic, proactive role in "striving for . . .higher-order goals and experienced meaningfulness associated with goal fulfillment." They argued that individual characteristics and higher-order goals interact to make performance at work meaningful. The authors cite the work of Weick on sensemaking (e.g., Weick, 1995, 2012), who explained how experiences at work are transformed into communicative practice as recursive social interaction. According to Barrick et al. (2013), "employees actively engage in an interpretive process to make meaning of their own jobs, roles, and selves at work by comprehending, understanding, and extrapolating cues received from others" (p. 147).

In other words, the subjectively experienced motivational state is a product, first, of the situation, but secondly, of how this situation is interpreted through social sense-making through language. This process should in turn affect the experienced levels of effort and quality exerted at work, together with a general sense of wellness, as experienced in the intention to stay in this job and as commitment to the organization. The chosen framework gives us the opportunity to operationalize situations using JCM and later extensions, predict ratings of motivations and outcomes building on SDT, and explore whether item responses reflect job characteristics, interpretive processes, or both. We want to emphasize here that our main concern is not with the theories of motivation itself, but with the contextually determined interpretation of Likert-scale items. The present theories are chosen for the way they allow exploration of contextual variables that influence text interpretation as well as motivational effects, hence the inclusion of self-rated levels of motivational outcomes.

Since job characteristics and types of motivation have been object of extensive research as quoted above, our focus is on the prospect of exploring the interpretive, semantic process involved to which we now turn.

Semantic Analysis

Work on natural language parsing in digital technologies has yielded a number of different techniques used with increasing frequency in social science. We will not review these in depth here, but concentrate on a brief description of latent semantic analysis, the technique used in the present study.

Latent semantic analysis (LSA) is a mathematical approach to assessing meaning in language, similar to how the brain determines meaning in words and expressions (Landauer and Dumais, 1997; Kintsch, 2001; Dennis et al., 2013). The general principle behind LSA is that the meaning of any given word (or series of words) is given by the contexts where this word is usually found. Just as children pick up the meaning of terms by noticing how they are applicable across different situations, LSA is a mathematical technique for determining the degree to which two expressions are interchangeable in a language.

Latent semantic analysis does this by establishing a semantic space from existing documents such as newspaper stories, journal articles and book fragments. In these semantic spaces, documents are used as *contexts* and the number of times any word appears in each context is entered in a word-by-document matrix. This matrix can be created out of a smaller number of texts, but the best results are typically obtained with semantic spaces containing millions of words in thousands of documents (Dumais et al., 1988; Landauer and Dumais, 1997; Gefen et al., 2017). From here, LSA transforms the sparse word-by-document matrix into three new matrices through singular value decomposition, a technique similar to principal component analysis (Günther et al., 2015; Gefen et al., 2017). Finally, researchers may project new texts of interest into these matrices to obtain a numerical estimate for the degree to which they are similar in meaning.

In a series of recent studies, LSA techniques have been used to explore a range of phenomena in survey statistics. Correlations between constructs have been explained as a result of semantic overlap (Nimon et al., 2016), as are the relationships between leadership behaviors and outcomes (Arnulf et al., 2014) and variable relationships in the technology acceptance model (Gefen and Larsen, 2017). In the same way, construct overlap (the so-called "jingle-jangle fallacy") was demonstrated and possibly empirically validated with the use of LSA (Larsen and Bong, 2016). The technique has also been applied to individual characteristics in responses, such as diagnosing psychopathology (Elvevag et al., 2017; Bååth et al., 2019), establishing personality patterns (Kjell et al., 2019), or predicting individual survey responses (Arnulf et al., 2018b).

One application that we will use here builds on a previous study of how semantically driven respondents are (Arnulf et al., 2018d). The argument in this approach is that strong semantic relationships between items will create higher correlations. An item with the wording "I like my job" will correlate highly with "I enjoy my work" simply because they share the same meaning and the LSA cosine for the two sentences are 0.73. Conversely, for two items to validly obtain different scores, they need to have dissimilar meanings. The LSA cosine for the items "I like my job" and "Customers are demanding" is −0.03, and they are not necessarily correlated even if they sometimes could be.

It is possible then to assess how similar any individual's set of scores is by calculating the distances between each pair of item scores. This approach has been investigated in four independent samples and was found to correspond to the response pattern predicted by LSA values (Arnulf et al., 2018d). Not all Likert scale instruments are equally semantically determined, and some seem entirely devoid of semantic predictability – the text algorithms may detect patterns but these do not seem to predict patterns in human responses (Arnulf et al., 2014). To the extent that a survey has a demonstrable semantic structure, we can assess the degree to which each single respondent is *compliant* with the semantic structure of the survey. To the degree that people are semantically compliant, they contribute to a response pattern that is semantically predictable, either as individuals or groups.

To compute semantic compliance, we first create a score distance matrix for each individual. The score distance matrix

is similar to the correlation matrix for the sample, but consists of the absolute difference in score level between two of the individual's scores [abs(score1-score2), abs(score1-score3)...]. We can then regress the individual's score distances on the semantically calculated matrix from LSA (Benichov et al., 2012; cf. Arnulf et al., 2018d). Take the three items used as example above: assume that to the items "I like my job," "I enjoy my work," and "Customers are demanding," our respondent answers 5, 5, and 2. The distance matrix between the three responses would be (5–5 = 0), (5–2 = 3), and (5–2 = 3). The series of LSA cosines 0.73, −0.03, and −0.03 are correlated −1.0 with the score differences (note the negative sign – *higher* overlap in meaning will result in *smaller* score distances).

As an operationalized measure of semantic compliance, we keep the unstandardized slope from the regression for each individual. If we regress the score distances above on the cosines, we get a slope of −3.95. The further from the semantically expected pattern (the weaker the slope), the more the individual may have made a personal interpretation of an item that departs from the semantically expected. We use this unstandardized slope as a measure and operationalization of how closely the single respondent matches a response pattern as predicted by the semantic algorithm alone.

Hypotheses About the Meaning of Motivational Items

Our unique approach to the measurement of motivation is now based on the combination of two approaches: examination of score levels and semantic compliance across a group of professions with different job characteristics. According to JCM, holders of jobs should display different motivational levels if the characteristics of the job also vary along the dimensions proposed by the theory. In other words, we are looking for response characteristics due to job types instead of individual differences (Hackman and Oldham, 1976; Chiu and Chen, 2005). However we are looking for two types of differences emanating from different job characteristics: The first would be the expected differences in motivational score levels, based on the influence that job characteristics are theoretically supposed have. The second is if different job characteristics will also influence the understanding of survey items in a way that is detectable by text algorithms.

This second type of differences goes back to Likert's (1932) original claim that verbal statements are beyond methodological reach. If we can begin to explore how different groups of respondents are systematically different in their response patterns, we can expand our tools of measurement beyond the simplification inherent in pure scale values. We can then begin to assess the impact of semantic factors such as context dependence, communities of practice, and social desirability, to name a few. By seeking a wide variation in possible job characteristics, we aimed to explore how semantics would explain the similarities and differences in frequently used measures of subjectively perceived motivation. Our exploration was guided by four hypotheses.

The first possibility we want to explore is if it is possible to show that reported levels of motivation are dependent on how the respondents interpret the items. If this is true, then the motivational levels will not only depend on the job type. The reported level of motivation will also depend on semantic compliance (i.e., differences in interpretation of items). Moreover, since different contexts will influence what the items mean to the respondents, these sources of variance will interact with each other. So, the main purpose of our study can be summed up in as follows:

H1: Self-reported levels of motivation differ by job type and the interaction between job type and semantic compliance.

However, the effects we look for in H1 are all taking place in the same responses – job holders who rate their levels of motivation are also displaying semantic characteristics. This risks a same-source bias, begging the question of which effect might be an artifact of the other (Podsakoff et al., 2012). We therefore want to follow the dynamics of semantics by tracing the effects of semantics to data sources independent of the subjective raters themselves. We start unpacking the problem by a series of hypotheses that relate to independent data. Our first independent data point is the salary level of each profession, not as self-rated but as the levels estimated by the national bureau of statistics in Norway (SSB). There are several reasons for choosing this type of data.

First, the salary levels of a profession in society is linked to the market value of this profession (Obermann and Velte, 2018). The mutual differences between salary levels of professions will be mixed a function of social status and macro-economic evaluation in the job markets, with possible effects on the interpretation of survey items. Secondly, research on JCM and on SDT (Kuvaas, 2006b; Deci et al., 2017) shows that monetary rewards have complicated effects on motivation its outcomes on work. Payment systems may exert a negative effect through perceptions of external control and counter-productive work focus. On the other hand, higher level of payment may signal recognition, status and power in ways that were predicted to increase IM in the theory of purposeful behavior (Barrick et al., 2013). We will therefore explore the extent to which semantic compliance relates to salary levels:

H2: By job type, semantic compliance of job type holders differ by salary levels.

In establishing the second independent rating, we look for the job characteristics as perceived by others. This is our second independent data point and replicates the original study of Hackman and Oldham (1976), who also used an external panel of raters to test JCM. A fundamental condition for influencing motivation by designing or crafting jobs is that there are some characteristics that will be apparent to most people, whether they hold the actual job or not. In the next hypothesis, we repeat this but look for differences in semantic compliance instead of motivational levels. On the other hand, the general public's perception of the characteristics and status of a job may in part be influenced by its market value, as indicated by salary levels. Our aim is to show that:

H3: By job type, external panel opinions of job characteristics differ by semantic compliance of job type holders, even when controlled by salary.

Finally, one may ask if these dynamics are of practical importance. If situational characteristics influence both the measurement values *and* the measurement instruments, one must expect that differences in motivational levels between groups may be evened out by the interpretative sense-making process (Barrick et al., 2013). People with different work contexts may make similar ratings of their motivational level. As noted by the authors of JCM and SDT, the general public perceives notable differences in job characteristics across society (Oldham and Hackman, 2010; Deci et al., 2017). We therefore expect a panel of raters to rate the job characteristics as more diverse than the job holders will rate their motivational levels:

H4: The standard deviation in the panel's job characteristics will show a greater dispersion of scores than the dispersion of self-rated motivational scores.

MATERIALS AND METHODS

The following sections describe the source of the data collected, measures used and analyses employed. Each is described in detail.

Data

The data used in this study represent four completely independent sources. We gathered self-reported levels of motivation from 399 respondents holding 18 different job types. In this context, we want to point out that we use the label "job type" as a simple descriptor of the work situations and characteristics that normally apply to holders of such jobs. Next, we obtained a panel of 30 persons rating the various job characteristics for each of the job types. The public income statistics were yet another dataset. Finally, the fourth dataset was made up of LSA semantic similarity indices computed on the item texts alone.

Participants

The original study of Hackman and Oldham (1976) claimed to survey a broad range of job characteristics, but the actual range of these characteristics was not described and seems as if their samples were from varying professions within the companies that participated in the survey. To test our hypotheses, we chose to aim for the broadest possible range of job characteristics within a society. Our self-report motivation sample therefore consisted of 399 persons from 18 job types. We aimed for equal sizes for ease of analysis, but this was difficult as the willingness to participate varied greatly across the job types. The number of 20 respondents in each group was chosen partly to balance the most reluctant groups of participants, and partly because groups of this size have previously been found to display consistent semantic behavior (Arnulf et al., 2018a,d; Arnulf and Larsen, 2020). We offer here a brief description of the job types and how respondents were enlisted:

Chief Executive Officers (CEOs) are very well paid, and wield much power. They responded willingly and our sample contains some of Norway's most high-profiled CEOs. As a contrast, we obtained a sample of street magazine vendors. These are generally drug addicts or other socially disadvantaged people who are given this job as a respectable means to make a living. They earn very little and only based on their sales. Others who earn little are a sample of volunteers from NGOs who enlist because of their support for a cause. Similarly ideologically inclined but also paid were a group of priests from the Church of Norway. As an assumed contrast to the purely value-based jobs, we enlisted a group of sex workers. This posed some difficulties as buying (but not selling) sex is illegal in Norway, leading to some reluctance in accepting contact. Some of the subjects were working in the streets and surveyed in a sheltering home, while others were contacted through online escort services. Another group was made up of purely professional soldiers, that is, who had been in paid combat service not as a part of mandatory military service or as part of a planned military career. Many of these did not want to give away their e-mail addresses, responding instead to paper and pencil versions of the survey. These groups were not easy to reach, but answered generously once they understood our request. We also contacted professions with high performance pressure such as professional athletes, artists, and stock brokers. The other groups could be seen as less extreme in job characteristics, such as car sales representatives, farmers, lawyers, morticians, dancers, and photographers. Taken together, we assumed that these groups would represent the true variation of motivationally relevant job characteristics in society. The cleaners and street magazine sellers were least willing to participate. The priests and the farmers were most enthusiastic and expressed happiness that someone was interested in their working conditions.

In total, we contacted 1,051 individuals as possible job holders but of these, only 504 potential respondents were identified to be in our target groups and asked to fill out a survey. Our 399 responses make up 79% of these 504 potential respondents. **Table 1** shows the 18 job types with the number of participants and gender distribution. Due to the sensitive nature of some professions, we refrained from asking about personal data from the respondents, but we did ask about gender even if this was not mandatory. Several groups appeared inclined to skip the gender question, resulting in large numbers of "unknown."

Panelists

Following the approach of Hackman and Oldham (1976) the job characteristics were rated by an external evaluation panel. The panel consisted of 30 individuals working in Norway with no relationships to the first sample or knowledge about the purpose of the study. The panel was recruited as a convenience sample from the researchers' own network. The inclusion criteria aimed simply to attain a representative group of adults with knowledge about the working world with dispersed demographics, resulting in 53% females with an age span of 17–62 years. The sample rated the job types on the JCM dimensions in order to obtain independent evaluations of perceived job characteristics associated with each job type. The panel members individually filled out a Norwegian-language web-based or paper survey.

TABLE 1 | Number of participants and distribution of gender for each job type.

Job type	Male	Female	Unknown	Total
Artist	10	12	0	22
Athlete	6	8	6	20
Bouncer	3	0	14	17
Car Sales Rep.	22	1	0	23
CEO	17	7	0	24
Cleaner	4	8	1	13
Dancer	2	10	8	20
Doctor	7	6	9	22
Farmer	8	4	27	39
Lawyer	13	7	0	20
Magazine Seller	13	5	0	18
Mortician	13	8	0	21
Photographer	10	11	0	21
Priest	23	11	4	38
Sex Worker	1	12	9	22
Soldier	10	6	3	19
Stockbroker	16	2	2	20
Volunteer	3	3	14	20
Total	181	121	97	399
% of Total	45.4	30.3	24.3	100

Income

Our source of information about income for the job types was the Norwegian National Statistics Bureau, SSB. These data were not collected from the respondents themselves, but consist entirely of the average income levels as listed by SSB in 2018.

Semantic Similarity Indices

The text of all the survey items was projected into a semantic space that we created out of texts from journal articles in the field of psychology. We termed this semantic space "psych" to denote its semantic heritage from psychological texts. This procedure returned a list of semantic cosines for ([50*49]/2) = 1,225 unique item pairs. This is the semantic equivalent of the correlation matrix (Arnulf et al., 2018d), and we will refer to this as LSA cosines or semantic similarity indices. The software for creating semantic spaces and projecting texts can be found as packages in Python (Anandarajan et al., 2019) or R (Günther et al., 2015; Wild, 2015; Gefen et al., 2017).

Semantic values raise a problem with negative correlations, because the cosines almost never take negative values. When they do, the negative sign can be read simply as very distant in the semantic matrices. Negative values do not indicate "opposite" as in correlations, where "like" is the opposite of "not like." In this study, we handled negative correlations by reverse-scoring all negatively worded items. This is often done with reversed items within scales. Additionally, to avoid the problem of negative cosines, we also reverse-scored two scales that are always negatively related to all the others, Turnover intention (TI) and economic exchange (EE).

Likert-Scale Measures

We will here describe in detail the self-rating scales on eight motivational constructs, along with the measurement instrument for job characteristics and the data on pay levels. Since motivation is a latent construct, we have chosen to include measures of motivational states together with their purported outcomes. A broader set of items allows a clearer analysis of semantic influences. Also, the inclusion of the outcomes lets us detect if the motivational effects vary along the motivational states as semantically predicted.

Self-Rated Motivation

We assembled a series of eight commonly used scales for measuring motivation in conjunction with self determination theory (SDT), totaling 50 items. All items were measured using a five-point Likert response scale ranging from 1 (strongly disagree) to 5 (strongly agree) and administered through a web- and paper-based survey. The first three variables – intrinsic motivation, with social and EE – can be seen as expressions of motivational states. The next four – citizenship behaviors, TI, work effort (WE) and work quality (WQ) – can be seen as outcome measures. The measures in the questionnaire are as follows.

Intrinsic motivation is defined as to "perform an activity for itself, in order to experience the pleasure and satisfaction inherent in the activity" (Kuvaas, 2006b, p. 369). This was assessed with a six-item scale developed by Cameron and Pierce (1994). One example item is 'My job is so interesting that it is a motivation in itself.'

Social exchange (SE) entails "unspecified obligations such that when an individual does another party a favor, there is an expectation of some future return. When the favor will be returned, and in what form, is often unclear" (Shore et al., 2006, p. 839). In contrast, EE involves transactions between parties that are not long-term or on-going but encompass the financial oriented interactions in a relationship. The constructs SE and EE were measured by a 16-item scale developed and validated by Shore et al. (2006) and previously used in a Norwegian context (Kuvaas and Dysvik, 2009). The SE and EE constructs were each measured with eight items. An example EE item is 'I do not care what my organization does for me in the long run, only what it does right now.' An example SE item is 'The things I do on the job today will benefit my standing in this organization in the long run.'

Organizational citizenship behavior (OCB) is defined as the "individual behavior that is discretionary, not directly or explicitly recognized by the formal reward system, and that in aggregate promotes the effective functioning of the organization" (Organ, 1988, p. 4). The construct was assessed with a seven-item measure validated by Van Dyne and LePine (1998). An example item is 'I volunteer to do things for my work group.'

Affective organizational commitment (AOC) can be defined as "an affective or emotional attachment to the organization such that the strongly committed individuals identifies with, is involved in, and enjoys membership in, the organization" (Meyer and Allen, 1997, p. 2). AOC was measured with six items previously used by Kuvaas (2006b), originally developed by Allen and Meyer (1990). A sample item is 'I really feel as if this organization's problems are my own.'

TABLE 2 | Job characteristic descriptions and items for the external evaluation panel.

Job Characteristic	Description	Question asked for each job type
Autonomy	"The degree to which the job provides substantial freedom, independence, and discretion to the individual in scheduling the work and in determining the procedures to be used in carrying it out" (Hackman and Oldham, 1975, p. 162).	The job gives a person considerable opportunity for independence and freedom in how he or she does the work.
Feedback	"The degree to which carrying out the work activities required by the job results in the individual obtaining direct and clear information about the effectiveness of his or her performance" (Hackman and Oldham, 1976, p. 258).	To what extent does doing the job itself, managers or co-workers or cooperation with others provide the person with information about his or her work performance?
Pay	Fixed regular payment an employee receives as a compensation for the employment.	Do you think this profession would be a nice profession if money had not been a problem?
Power	"Absolute capacity of an individual agent to influence the behavior or attitudes of one or more designated target persons at a give point in time" (Yukl, 2012b, p. 189).	Do you think this profession implies the ability to execute power?
Prestige	"By educational attainment, by occupational standing, by social class, by income (or poverty), by wealth, by tangible possession" (Hauser and Warren, 2012).	I would have bragged about having this profession to others.
Relatedness	"Both experiencing others as responsive and sensitive and being able to be responsive and sensitive to them – that is, feeling connected and involved with others and having a sense of belonging" (Ryan and Deci, 2017, p. 86).	Do you think this profession contains meaningful relationships with other people?
Safety/danger	Risks of being injured at work.	Do you think this profession is exposed to any risk/danger?
Skill variety	"Degree to which a job requires a variety of different activities in carrying out the work, involving the use of different skills and talents of the employee" (Hackman and Oldham, 1975, p. 161).	How much variety is there in the job? That is, to what extent does the job require a person to do many different things at work, using a variety of his or her skills and talents?
Task identity	"The degree to which the job requires completion of a 'whole' and identifiable piece of work; that is, doing a job from beginning to end with a visible outcome" (Hackman and Oldham, 1975, p. 162).	The job provides a person with the chance to finish completely any work he or she starts.
Task significance	"The degree to which the job has a substantial impact on the lives or work of other people, whether in the immediate organization or in the external environment" (Hackman and Oldham, 1975, p. 161).	In general, how significant or important is the job? That is, are the results of the person's work likely to significantly affect the lives or well-being of other people?
Work-life balance	"An individual's ability to meet their work and family commitments" (Delecta, 2011, p. 187).	Do you think this profession enables a person to balance work and leisure?

Turnover intention may be defined as "behavioral intent to leave an organization" (Kuvaas, 2006a, p. 509). The five items were retrieved from Kuvaas (2006a). One example item is 'I will probably look for a new job in the next year'.

Work quality is defined as "quality of the output" (Dysvik and Kuvaas, 2011, p. 371), while WE is defined as "the amount of energy an individual put into his/her job" (Buch et al., 2012, p. 726). Kuvaas and Dysvik (2009) developed a scale with five items for each. A sample WE item is 'I often expend extra effort in carrying out my job,' while a sample WQ item is 'I rarely complete a task before I know that the quality meets high standards.'

Job Characteristics Model (JCM)

Eleven different characteristics connected to JCM were identified and operationalized as single items for each job type, and rated by our panel (see **Table 2**). The items for autonomy, feedback, skill variety, task identity and task significance were developed by Hackman and Oldham (1975) as part of their original research. As outlined by Barrick et al. (2013), and also as indicated by a later review of JCM (Oldham and Hackman, 2010), there are more characteristics that may activate motivational states than what was originally assumed, particularly related to prestige, power, and other social characteristics. We therefore asked the panel to also rate the jobs on work-life balance, power, safety/danger, prestige, and relatedness (Delecta, 2011; Hauser and Warren, 2012; Ryan and Deci, 2017). To avoid a cumbersome number of items for the panel to fill out, we followed the original procedure from JCM using single-item questions about characteristics for each profession (Hackman and Oldham, 1976).

Analyses

We began our analyses by computing semantic compliance so that we could build our participant database. Semantic compliance (or similarity with the semantic matrix) was created for each participant by regressing the absolute difference between item scores (i.e., individual item distance matrix) on corresponding LSA cosines that were derived from the psych semantic space (i.e., semantic similarity matrix) and saving the *unstandardized slope* (Benichov et al., 2012; cf. Arnulf et al., 2018d).

A series of regression analyses were conducted to determine to what extent job type and the interaction between job type and semantic compliance explained the variance in motivation scores, thereby allowing us to simultaneously look at differences between and within job type as predicted in H1. To interpret the regression effects, we used regression commonality analysis (cf. Nimon et al., 2008). We then aggregated self-reported levels of motivation and external panel opinions of job characteristics by job type, and

explored first how salary levels predicted semantic compliance (H2), next how job characteristics as rated by the external panel predicted semantic compliance (H3), and finally if the dispersion of scores was different in the panel and self-rating groups (H4).

RESULTS

We first present the overall score levels and relationships for the participant data (see **Table 3**) before proceeding to the hypotheses analyses. Across all job types, semantic compliance had a mean of -0.16 (SD = 0.4). This implies that on average, participants showed a tendency to be semantically compliant. Further, semantic compliance was most highly related to score levels on TI, affective commitment (AC), WQ, and IM. Note that TI and EE are reverse-scored. The alpha coefficients of all scales were generally high (0.75 − 0.90) and they generally correlate quite highly with each other. In particular, TI tends to correlate highly with all other scales, while WQ usually displays the lowest correlations with other scales.

Hypothesis 1

Hypothesis 1 considered whether *self-reported levels of motivation differed by job type and the interaction between job type and semantic compliance*. To test H1, we ran regression analyses on each eight motivational scales using job type and the interaction between job type and semantic compliance as predictors. The results can be seen in **Table 4**. Across most motivational scales, job type and the interaction between job type and semantic compliance contributed significantly to the explained variance, supporting H1. While job type alone mostly has a greater explanatory effect on most score levels than the interaction between job type and semantic compliance, this relationship varies visibly across the scales. In the case of TI, the interaction between job type and semantic compliance predicts motivational level better than job type.

Using the regression results, we also looked at whether respondents with high, average or low semantic compliance had significantly different score levels on each scale (see **Figures 1–8**). It appears that some groups display more semantic disparities than others, and some scales also create greater differences within job types than others. Interestingly, each profession differentiated in the association between their semantic compliance and self-reported levels of motivation for at least one measure.

The largest differentiation in semantic compliance takes place in responding to TI. Eight job types display significant differences in score levels based on their semantic compliance: athletes, bouncers, dancers, doctors, lawyers, magazine sellers, soldiers and stockbrokers. Next, for AC, there are five groups displaying significant differences in score level depending on semantics: artists, bouncers, doctors, magazine sellers, and morticians.

Conversely, some scales do not seem to elicit much within-group differences. For WE, only priests seem to differentiate. For EE, only bouncers and CEOs differentiate, and for IM, only bouncers and magazine sellers do.

The box plot for turnover intention also shows a general trend for the whole sample, namely, that higher semantic compliance is often related to somewhat lower or at least moderated mean score levels (note that turnover intention as a scale is reverse-scored in our analysis). There are only two notable differences, volunteers and sex workers, whose values are not significantly different from zero.

Two interesting cases are WQ and WE. These are the scales where the differences between groups are least pronounced. There are still discernible within-group differences in score levels and semantic compliance, enough to make high scorers less semantically compliant. In the case of WQ, where all groups score about the same, semantics explain almost as much unique variance as the score level differences (35% vs. 49% of the explained variance).

Together, **Table 4** and the box plots in **Figures 1–8** show that different job types will have different impacts on the relationship between semantics and score levels. There is no single, simple relationship between the two. Instead, the same groups of items seem to be interpreted so differently within and between groups that there will be significant differences in score levels depending on these differences. Looking at the relationship between semantics and motivational scales, a pattern emerges that may be due to semantic uncertainty where respondents differ.

Even if the interactions are complex, there are also some more linear relationships between semantics and motivational levels. **Table 5** sorts mean self-reported levels of motivation from least to most semantically compliant. Aggregated by job type, the mean motivational measures of turnover intention and OCB were the most semantically related but in opposite directions and the mean motivational measures of economic exchange and WE were the least semantically related (see **Table 6**). Taken together, these findings support H1.

Hypothesis 2 and 3

Hypothesis 2 and 3 examined data aggregated by job type and considered whether *salary levels* (H2) and *external panel opinions of job characteristics controlled by salary levels* (H3) *differed by semantic compliance of job type holders*. Interestingly,

TABLE 3 | Correlation matrix and descriptive statistics for semantic compliance and self-reported levels of motivation.

	SC	AC	EE[a]	IM	OCB	SE	TI[a]	WE	WQ
AC	0.26	0.75							
EE	0.03	0.52	0.84						
IM	0.14	0.59	0.55	0.90					
OCB	−0.01	0.34	0.26	0.26	0.87				
SE	−0.03	0.47	0.28	0.45	0.35	0.80			
TI	0.38	0.50	0.48	0.62	0.13	0.41	0.89		
WE	0.11	0.37	0.28	0.53	0.39	0.31	0.31	0.78	
WQ	0.15	0.13	0.05	0.29	0.38	0.21	0.16	0.53	0.75
M	−0.16	3.73	3.96	4.30	3.95	3.71	4.04	4.26	3.92
SD	0.40	0.79	0.83	0.76	0.65	0.71	1.02	0.60	0.54

[a]*Reverse coded. Coefficient alpha along the diagonal. SC, semantic compliance. AC, affective commitment. EE, economic exchange. IM, intrinsic motivation. OCB, organizational citizenship behavior. SE, social exchange. TI, turnover intention. WE, work effort. WQ, work quality.*

TABLE 4 | Regression results for motivation measures by job type (JT) and the interaction of job type and semantic compliance (SC).

	AC			EE[a]			IM			OCB			SE			TI[a]			WE			WQ		
Job Type	b_0	b_1	p	b_0	b_1	p	b_0	b_1	p	b_0	b_1	p	b_0	b_1	p	b_0	b_1	p	b_0	b_1	p	b_0	b_1	p
Artist	3.46	1.88	**<0.01**	4.24	0.31	0.55	4.86	0.27	0.59	3.15	1.25	**0.01**	3.40	0.95	0.06	4.50	0.75	0.25	4.62	0.22	0.61	4.31	0.50	0.21
Athlete	3.77	0.21	0.51	3.90	0.08	0.82	4.43	0.12	0.69	3.97	−0.15	0.61	3.62	−0.40	0.21	3.91	1.61	**<0.01**	4.48	0.16	0.57	3.94	−0.01	0.97
Bouncer	3.32	1.31	**<0.01**	3.46	0.99	**<0.01**	3.95	1.46	**<0.01**	4.17	−0.01	0.95	3.41	1.10	**<0.01**	3.43	1.75	**<0.01**	4.25	−0.05	0.81	3.99	−0.25	0.23
Car Sales Rep	3.79	0.41	0.26	3.93	−0.15	0.68	4.36	0.36	0.30	4.38	−0.27	0.41	4.12	−0.39	0.28	4.17	0.77	0.09	4.44	−0.08	0.80	4.01	0.66	**0.02**
CEO	4.16	−0.09	0.79	4.32	−0.71	**0.04**	4.71	−0.21	0.52	4.30	−0.22	0.48	4.06	−0.48	0.16	4.66	0.35	0.42	4.45	0.18	0.54	3.90	0.02	0.94
Cleaner	2.76	0.32	0.42	2.97	−0.24	0.55	2.91	−0.45	0.23	3.77	−0.59	0.09	3.30	−0.72	0.06	2.89	0.61	0.22	3.73	−0.09	0.78	3.78	−0.50	0.10
Dancer	3.99	0.39	0.38	4.07	−0.51	0.26	4.47	0.18	0.66	3.81	0.08	0.83	3.73	−0.05	0.91	3.85	1.11	**0.05**	4.33	0.37	0.32	3.84	0.67	**0.05**
Doctor	3.75	0.72	0.05	4.38	0.53	0.15	4.51	0.17	0.62	3.98	0.34	0.29	3.72	−0.21	0.55	4.40	1.19	**0.01**	4.23	0.50	0.10	3.79	0.20	0.48
Farmer	3.99	0.26	0.35	4.32	0.03	0.90	4.40	0.02	0.93	3.96	−0.17	0.51	3.71	−0.50	0.07	4.40	0.13	0.71	4.26	0.22	0.35	3.78	0.11	0.63
Lawyer	3.38	−0.50	0.24	4.12	0.33	0.44	3.93	−0.26	0.52	3.97	−0.39	0.30	3.86	−0.10	0.81	3.92	1.68	**<0.01**	4.26	0.08	0.83	3.91	−0.02	0.95
Magazine Seller	3.20	1.32	**<0.01**	2.92	0.39	0.29	3.63	0.84	**0.01**	3.33	−0.05	0.88	3.27	0.74	**0.04**	3.37	1.59	**<0.01**	3.92	0.26	0.39	3.64	0.01	0.96
Mortician	3.93	0.64	**0.04**	4.26	−0.42	0.19	4.43	0.12	0.68	4.30	0.19	0.50	4.11	0.14	0.66	4.39	0.61	0.13	4.50	0.21	0.43	4.22	0.58	**0.02**
Photographer	4.37	0.60	0.22	4.34	0.21	0.68	4.77	0.10	0.83	3.89	0.88	**0.04**	4.14	−0.65	0.19	4.58	0.18	0.77	4.58	0.40	0.34	4.19	0.58	0.13
Priest	4.06	0.30	0.33	4.41	−0.01	0.98	4.60	0.19	0.52	3.94	0.42	0.13	3.60	−0.35	0.25	4.33	0.70	0.07	4.01	0.50	**0.05**	3.70	0.33	0.17
Sex Worker	3.21	−0.06	0.86	3.12	−0.34	0.31	3.85	−0.19	0.55	3.66	−0.11	0.71	3.49	−0.16	0.63	3.66	−0.29	0.48	3.64	−0.16	0.58	3.97	0.08	0.76
Soldier	3.64	0.43	0.15	4.09	−0.08	0.79	4.26	0.33	0.25	4.33	0.25	0.35	3.54	−0.31	0.29	3.69	1.12	**<0.01**	4.30	0.16	0.53	4.01	0.55	**0.02**
Stockbroker	3.20	0.73	0.08	3.04	0.14	0.74	3.89	0.19	0.63	3.59	0.18	0.62	3.65	0.53	0.19	3.61	1.89	**<0.01**	4.22	−0.06	0.86	3.82	0.09	0.77
Volunteer	3.80	−0.82	0.09	4.26	−0.16	0.75	4.34	−0.59	0.21	4.29	−0.20	0.64	3.73	−0.92	0.06	4.25	−0.67	0.29	4.35	0.14	0.74	3.89	0.49	0.20
F(35,363)			6.15			7.21			6.69			3.81			3.13			6.88			2.81			2.56
R^2			0.37			0.41			0.39			0.27			0.23			0.40			0.21			0.20
CC_{JT}			0.23			0.35			0.28			0.23			0.13			0.18			0.18			0.10
$CC_{JT:SC}$			0.13			0.04			0.08			0.05			0.10			0.18			0.03			0.07
$CC_{JT,\ JT:SC}$			0.02			0.02			0.03			−0.01			0.01			0.04			0.01			0.03

[a]*Reverse coded. AC, affective commitment. EE, economic exchange. IM, intrinsic motivation. OCB, organizational citizenship behaviors. SE, social exchange. TI, turnover intention. WE, work effort. WQ, work quality. $P \leq 0.05$ shown in bold. CC, commonality coefficient. In some cases, $\sum CC \neq R^2$ due to rounding errors.*

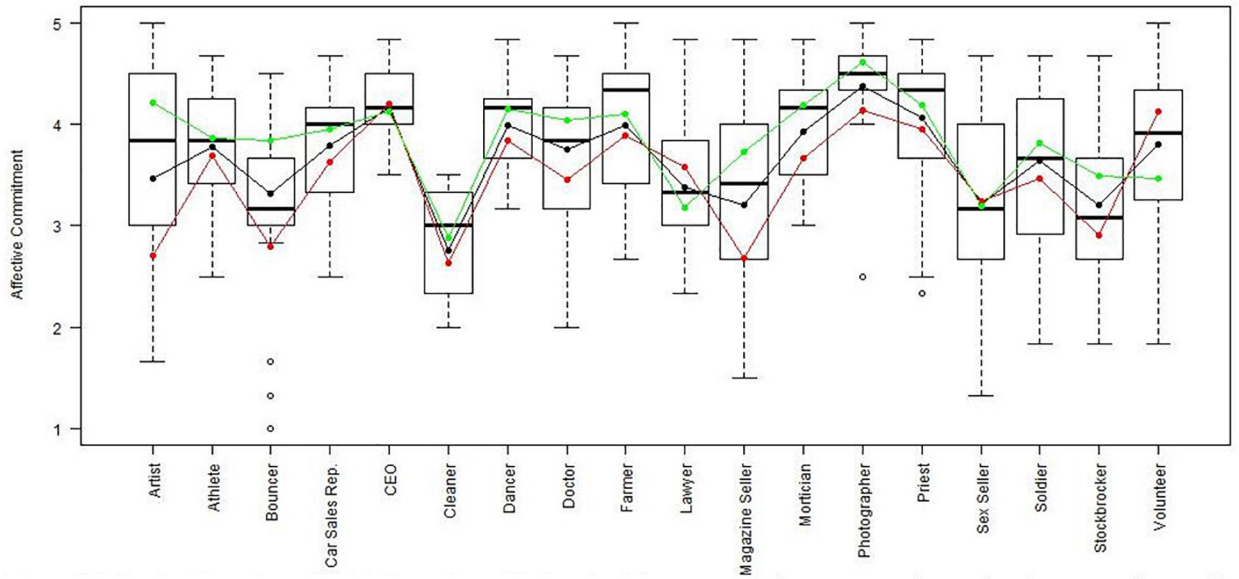


FIGURE 1 | Affective commitment by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD*), –0.16 (*M*), and –56 (M – 1 *SD*).

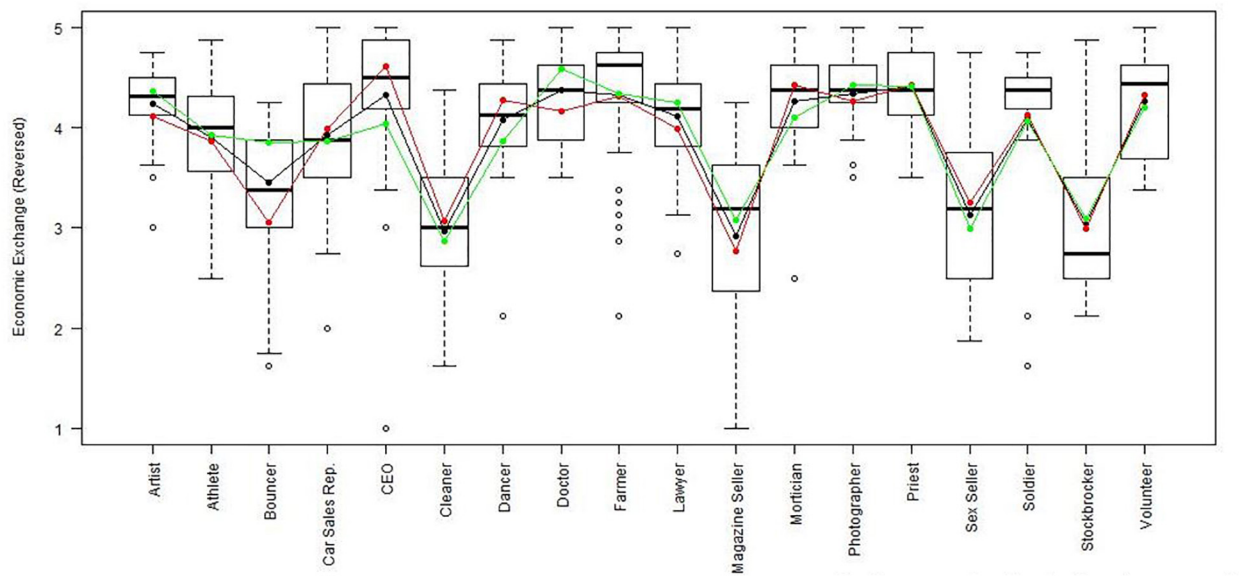


FIGURE 2 | Economic exchange (Reversed) by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD*), –0.16 *(M),* and –0.56 (*M* – 1 *SD*).

there are significant relationships between the four independent sources – national salary levels, panel-rated characteristics, self-rated motivation and semantic values. Group means for semantic compliance, salary, and the panel-rated characteristics are listed in **Table 7**, together with the inter-rater reliabilities of the panel characteristics ratings. The ICCs of the panel ratings are all above 0.92 except for the variable *task identity*, which is only 0.52. Salary turns out to be significantly related to semantic compliance of the job holders, as the rank-order correlation between semantic compliance and *salary* is −0.63. This supports H2. **Table 7** also shows a tendency for groups of high and low scores to cluster along the continuum made up by semantic compliance and income.

Table 6 shows how the panel's ratings of job characteristics show strong and significant correlations between job characteristics and motivational levels. In particular, the variables *autonomy*, *feedback*, and *skill variety* were strongly related to motivational variables in the direction suggested by

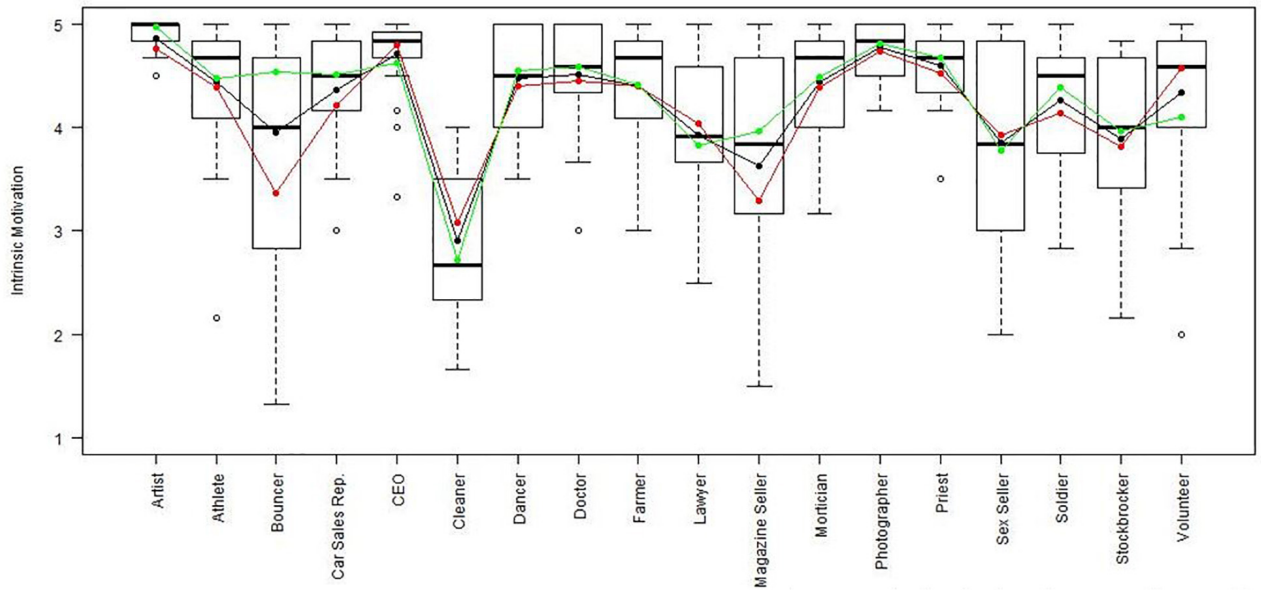


FIGURE 3 | Intrinsic motivation by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD*), –0.16 (*M*), and –0.56 (*M* – 1 *SD*).

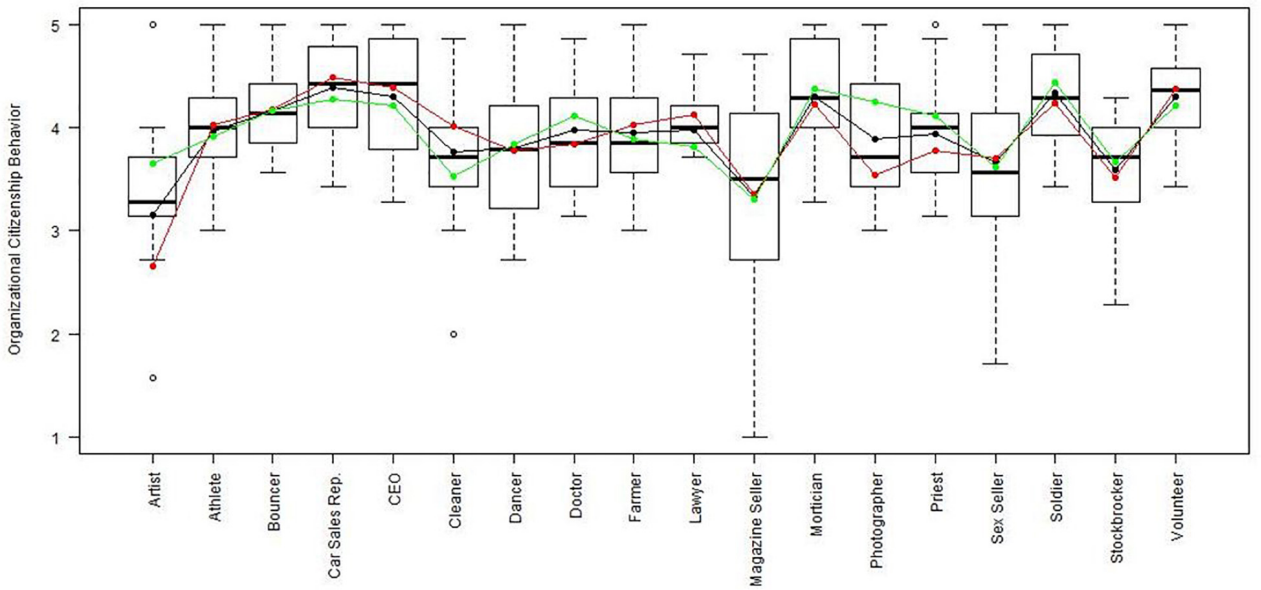


FIGURE 4 | Organizational citizenship behavior by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD*), –0.16 (*M*), and –0.56 (*M* – 1 *SD*).

JCM and SDT. Concomitantly, the variable "economic exchange" also correlates highly with the same variables.

Testing H3 raises an issue about sample size. The numbers are based on two samples – one with a panel of 30, the other with 399 respondents – but aggregated by job types the sample size is reduced to 18. The most conservative approach would be to look at relationships with a *p*-level above 0.05, $n = 18$. We find strong correlations between salary levels and the panel's perception of *power, prestige*, *feedback, worklife balance, safety/danger, skill variety*, and *task significance* ($|\rho| \geq 0.47$, $p \leq 0.05$, **Table 6**, rightmost columns). Only power and safety/danger as panel rated characteristics appear significantly related to semantic compliance. Controlling for salary, the only significant correlation between job characteristics and semantic compliance is safety/danger. However, considering that the numbers stem from bigger samples, there are sizeable

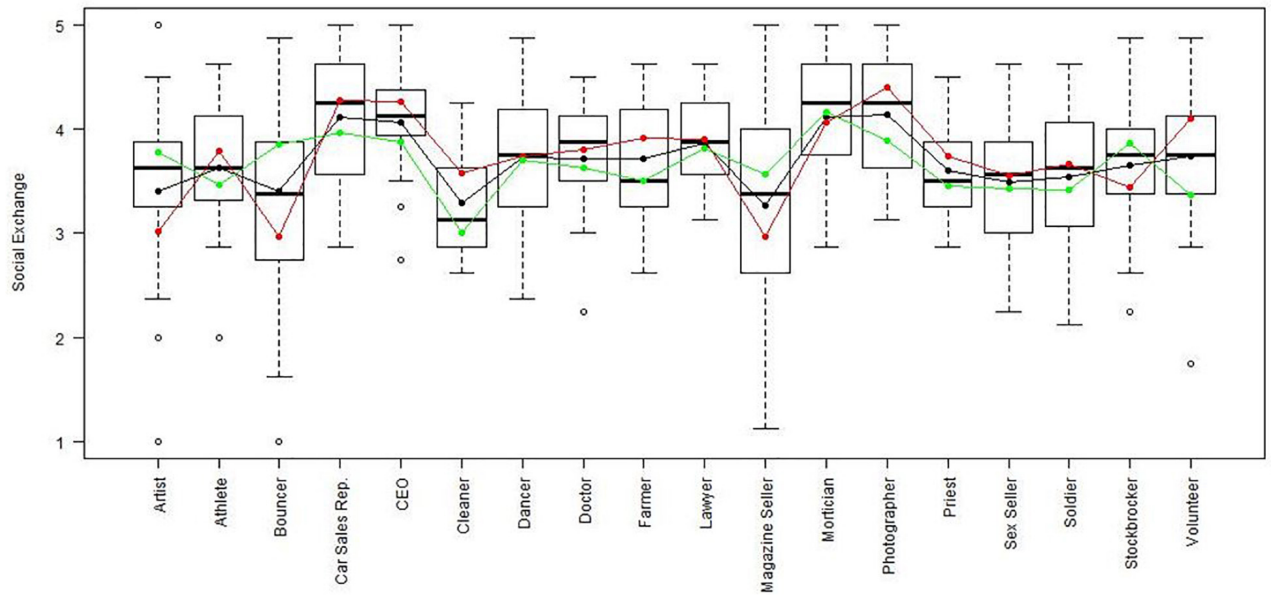


FIGURE 5 | Social exchange by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD),* –0.16 (*M*), and –0.56 (*M–1 SD*).

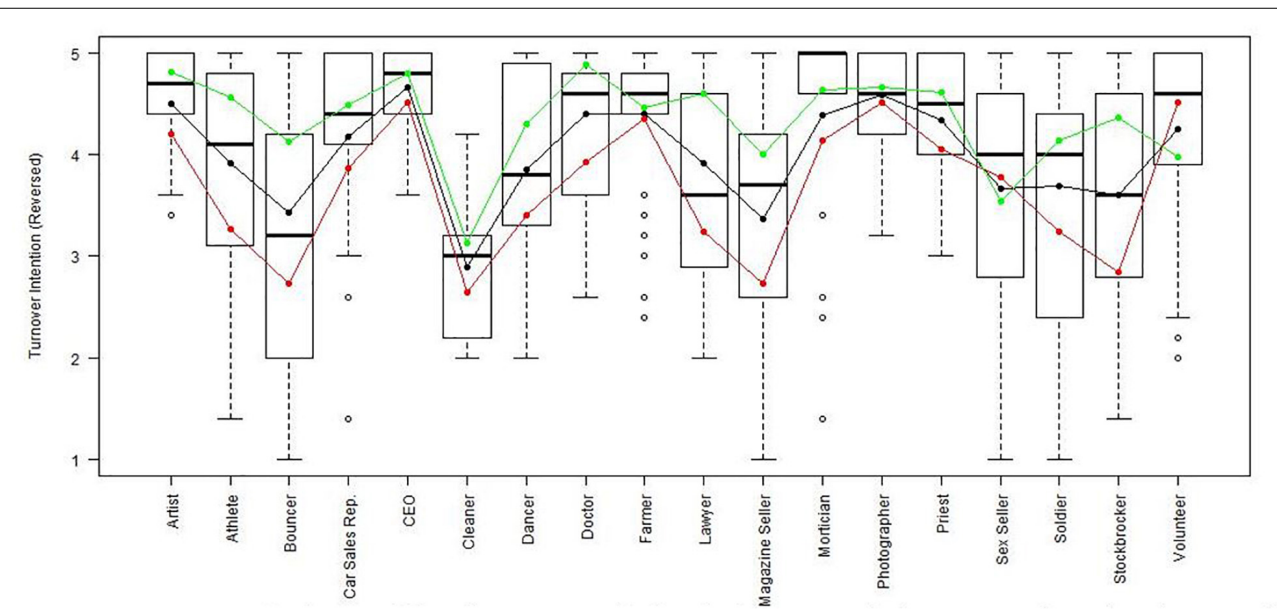


FIGURE 6 | Turnover intention (Reversed) by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD*), –0.16 (M), and –0.56 (*M* – 1 *SD*).

correlations with practical significance. Characteristics originally theorized to predict motivational levels, such as autonomy, feedback, power, relatedness, skill variety and task identity show medium to strong correlations with semantic compliance even after controlling for salary. The lowermost rows in **Table 6** show how semantic compliance correlates with the motivational scales themselves (from which the semantic compliance numbers are derived). These numbers are actually significantly lower than the correlations with the panel data ($p = 0.02$, Mann–Whitney test). H3 is therefore at least partly supported.

Hypothesis 4

The range of average scores on the motivational scales in **Table 5** is remarkably narrow. As argued in STSR, a score on a Likert item is en endorsement of a statement, in our case a motivational self-description. If we round the average scores to the nearest

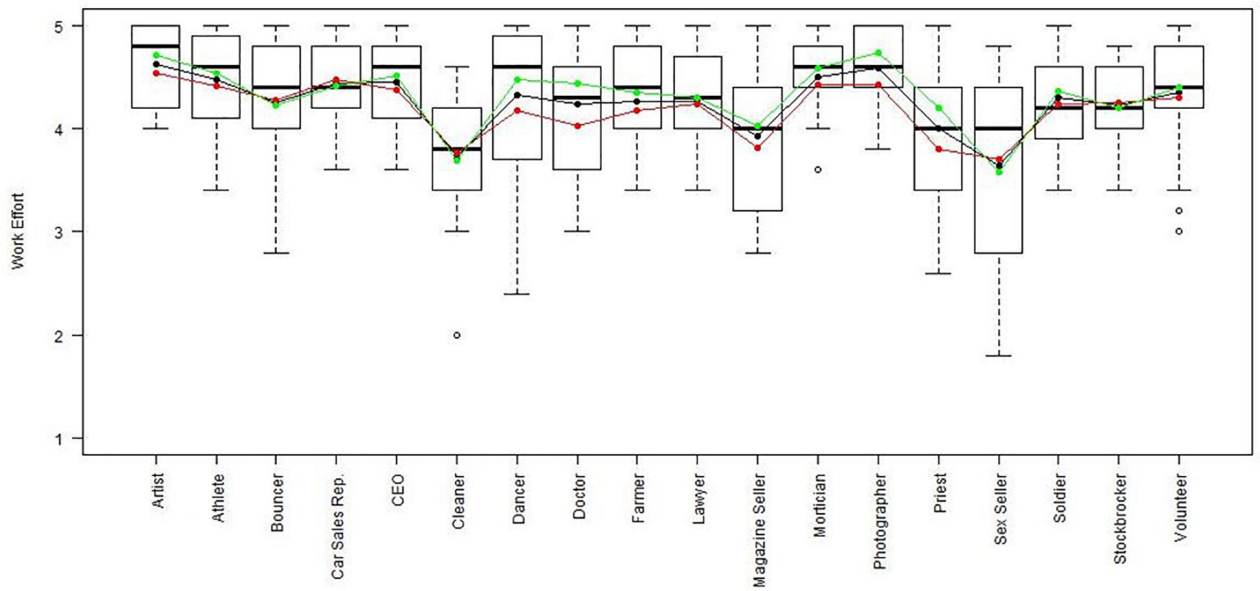


FIGURE 7 | Work effort by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD*), –0.16 (*M*), and –0.56 (*M* – 1 *SD*).

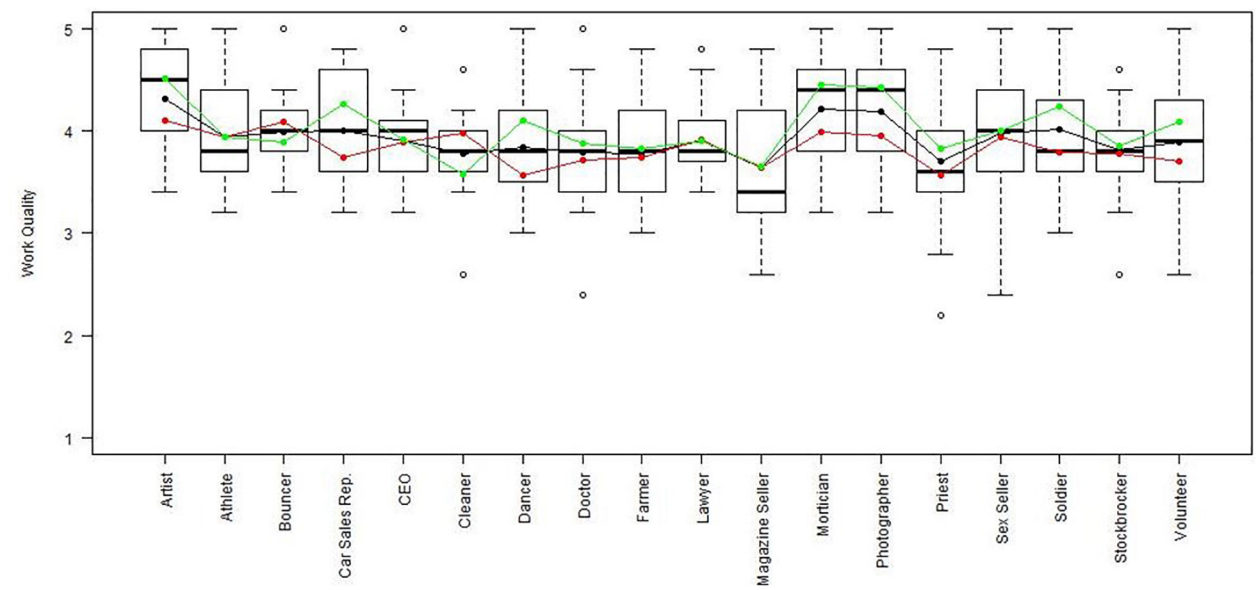


FIGURE 8 | Work quality by job type. Green, black, and red lines respectively represent estimates based on semantic compliance of 0.24 (*M* + 1 *SD*), –0.16 (*M*), and –0.56 (*M* – 1 *SD*).

integer and replace the integer with the corresponding statement on a motivational scale, the job types would literally describe their motivation in almost the same terms. The differences across job types within each scale exceeds 1 point in only two cases (IM and TI), where the differences do not exceed 2 points. H4 stated that *the standard deviation in the panel's job characteristics will show a greater dispersion of scores than the dispersion of self-rated motivational scores*. To test this we computed the standard deviation in the panel's rating of each characteristics across the job types. We then compare this to its counterpart in the self-rated group, by computing the standard deviation of mean scores across motivational levels and job types. The two sets of numbers are displayed at the bottom of **Tables 5**, **7**. It turns out that the variation in the panel's rating of job characteristics (0.83) is much higher than the variation in self rated motivational levels (0.38, $p = 0.001$ in a Mann–Whitney test), supporting H4.

DISCUSSION

The purpose of this study was to explore how different professional contexts influence the semantic patterns of responses to motivational items with ensuing consequences for score levels. Our findings supported the predictions from job design theory that levels of motivation differ significantly between job types according to their characteristics (Hackman and Oldham, 1975, 1976), but interestingly, the semantic characteristics of respondents also explained a substantial proportion of the differences in score levels. For most motivation measures, the interaction between job type and semantic compliance explained a substantive amount of unique variance in score levels, supporting H1. This suggests that scholars and scholar-practitioners may be mis-estimating the effect of job type on motivation when using traditional methods that do not consider participants' tendency to respond semantically.

Our findings imply that respondents from different job types differ substantially in how they perceive and interpret the items. Different job types do not only give people different subjective levels of motivation, but these job types also influence and probably change the meaning of each item. The effect is not a general methodological effect with equal impact across conditions, because some situations seem to alter the meaning of some scales more than others. This demonstrates that the relationship between job characteristics and self-rated motivation is not a two-way relationship. Instead, it is a three-way relationship, depending also on the subjects' semantic parsing of the items, which will vary systematically both between and within job types. Our finding is in line with the theory of purposeful behavior, which states that job holders will engage in sense-making activities to proactively create meaning in their situations (Barrick et al., 2013).

Since semantics and score levels are practically intertwined and difficult to separate (Arnulf et al., 2018d), the relationship between the two could possibly be interpreted as a methodological artifact such as common method variance (Podsakoff et al., 2012) or endogeneity (Antonakis et al., 2010). For that reason, we introduced two more independent data sources, an external panel and national statistics on salary levels. Interestingly, there was a strong correlation between the salary levels of the job types and the tendency of the job holders to respond semantically compliant.

This probably has several implications. One obvious reason for this finding is that the language in the survey items is most appropriate for people with high income. Another related reason is that high income is correlated with high social status and education, along with the linguistic habits and competence that come from such demographic variables. Among the most semantically predictable groups are highly trained academics such as lawyers and doctors, and athletes who tend to be competitively oriented and intellectually acute (Cooper, 1969). On the other side of the scale, the cleaners in our study had mostly either little education, or many of them were foreigners with high likelihood of lower language skills. One notable exception in the sample was the bouncers, who are not high earners but who scored very high on semantic compliance. This is a group of people who may be trained in using their verbal skills to deal with

TABLE 5 | Job type self-reported levels of motivation sorted by similarity compliance (SC).

		Motivational Measures							
Job type	**SC**	**AC**	**EE[a]**	**IM**	**OCB**	**SE**	**TI[a]**	**WE**	**WQ**
Artist	−0.04	3.68	4.27	4.89	3.30	3.51	4.59	4.65	4.36
Mortician	−0.06	3.99	4.22	4.44	4.32	4.13	4.45	4.52	4.28
Mag. Seller	−0.07	3.32	2.96	3.70	3.33	3.33	3.51	3.94	3.64
Farmer	−0.07	4.01	4.33	4.41	3.94	3.67	4.42	4.28	3.79
Car Sales Rep.	−0.08	3.83	3.91	4.39	4.36	4.09	4.23	4.43	4.05
Cleaner	−0.08	2.78	2.95	2.87	3.73	3.24	2.94	3.72	3.74
Photographer	−0.11	4.40	4.35	4.78	3.94	4.11	4.59	4.60	4.22
Priest	−0.12	4.07	4.41	4.61	3.96	3.59	4.36	4.02	3.71
Volunteer	−0.15	3.79	4.26	4.33	4.29	3.73	4.24	4.35	3.90
Dancer	−0.15	3.99	4.07	4.48	3.81	3.73	3.86	4.33	3.84
Sex Worker	−0.16	3.21	3.13	3.85	3.66	3.49	3.66	3.64	3.97
Stockbroker	−0.16	3.20	3.04	3.89	3.59	3.65	3.60	4.22	3.82
CEO	−0.18	4.17	4.33	4.72	4.30	4.07	4.65	4.44	3.90
Athlete	−0.27	3.75	3.89	4.42	3.99	3.67	3.73	4.46	3.94
Doctor	−0.31	3.64	4.30	4.48	3.93	3.75	4.22	4.15	3.76
Soldier	−0.31	3.57	4.11	4.21	4.29	3.59	3.52	4.27	3.93
Lawyer	−0.32	3.46	4.06	3.98	4.04	3.88	3.64	4.25	3.91
Bouncer	−0.33	3.10	3.29	3.70	4.17	3.22	3.13	4.26	4.04
SD		0.42	0.54	0.49	0.33	0.28	0.52	0.28	0.20

[a]*Reverse coded. AC, affective commitment. EE, economic exchange. IM, intrinsic motivation. OCB, organizational citizenship behaviors. SE, social exchange. TI, turnover intention. WE, work effort. WQ, work quality. SC, semantic compliance. Top six least semantic italicized and underlined. Top five most semantic bolded.*

TABLE 6 | Correlations between self-reported levels of motivation, semantic compliance, salary, and panel responses of job characteristics aggregated by job type.

	Motivational measures											
Measures	**AC**	**EE[a]**	**IM**	**OCB**	**SE**	**TI[a]**	**WE**	**WQ**	**SC**	ρ_{SC}	ρ_{Salary}	$\rho_{SC.Salary}$
Job characteristic												
Autonomy[b]	0.61	0.53	0.65	−0.24	0.30	0.70	0.45	0.21	0.33	0.30	0.04	0.42
Feedback[b]	0.46	0.56	0.56	0.38	0.66	0.36	0.59	0.21	−0.40	−0.48	0.49	−0.25
Work without pay	0.59	0.70	0.65	0.13	0.46	0.58	0.63	0.16	−0.13	−0.09	0.23	0.08
Power	0.21	0.45	0.36	0.34	0.31	0.23	0.32	−0.01	−0.64	−0.64	0.65	−0.38
Prestige	0.43	0.56	0.51	0.19	0.42	0.37	0.53	−0.06	−0.38	−0.40	0.50	−0.12
Relatedness	0.50	0.65	0.48	0.31	0.41	0.49	0.18	−0.08	−0.15	−0.03	0.33	0.24
Safety/danger	−0.25	−0.22	−0.15	0.01	−0.34	−0.36	−0.35	−0.21	−0.56	−0.62	0.46	−0.48
Skill variety[b]	0.46	0.64	0.57	0.20	0.40	0.50	0.29	0.01	−0.35	−0.34	0.71	0.21
Task identity[b]	0.18	0.04	0.10	−0.22	0.30	0.26	0.22	0.43	0.42	0.38	−0.21	0.33
Task significance[b]	0.28	0.61	0.32	0.38	0.27	0.35	0.16	−0.12	−0.36	−0.24	0.47	0.08
Worklife balance	0.02	0.04	−0.02	0.03	−0.13	0.11	0.12	0.32	0.43	0.44	−0.70	<0.01
SC	0.23	−0.02	0.11	−0.34	0.12	0.40	0.04	0.18	1.00	1.00	−0.63	
ρ_{SC}	0.31	0.11	0.24	−0.25	0.05	0.42	0.25	0.06				
ρ_{Salary}	0.13	0.32	0.12	0.24	0.22	0.10	−0.21	−0.17				
$\rho_{SC.Salary}$	0.52	0.43	0.41	−0.13	0.25	0.62	0.16	−0.06				

Unless otherwise noted, correlations are Pearson's r. [a]Reverse coded. AC, affective commitment. EE, economic exchange. IM, intrinsic motivation. OCB, organizational citizenship behavior. SE, social exchange. TI, turnover intention. WE, work effort. WQ, work quality. SC, semantic compliance. [b]Characteristics associated with Hackman and Oldham (1975) job characteristic model. Given a sample size of 18, absolute correlation coefficients of 0.71 are statistically significant at alpha = 0.001; 0.59 at alpha = 0.01, 0.47 at alpha = 0.05, and 0.40 at alpha = 0.10.

TABLE 7 | Job type panel responses of job characteristics and salary sorted by similarity compliance (SC).

		Job Characteristics											
Job Type	***SC***	**AU**	**FB**	**WPay**	**PWR**	**PR**	**RL**	**RISK**	**SV**	**TI**	**TS**	**WLB**	**Salary**
Artist	*−0.04*	**4.53**	3.47	**3.67**	2.67	3.13	2.93	*1.73*	3.43	**4.03**	3.07	**4.17**	*7,416*
Mortician	*−0.06*	*2.37*	3.73	*2.17*	*2.13*	*2.23*	**4.30**	*1.63*	*2.27*	**4.23**	**4.00**	3.70	*40,200*
Magazine Seller	*−0.07*	3.20	*2.33*	*1.83*	*1.20*	*2.00*	3.33	**3.10**	*2.03*	*3.60*	*2.43*	3.53	*1,000*
Farmer	*−0.07*	**4.10**	*2.60*	3.07	2.20	3.13	*2.60*	2.83	3.57	4.00	3.37	2.60	46,173
Car Sales Rep.	*−0.08*	*2.47*	3.67	*2.00*	*1.83*	*1.90*	*2.40*	*1.77*	*2.00*	3.77	*2.10*	**4.23**	*36,275*
Cleaner	*−0.08*	*1.97*	*2.57*	*1.47*	*1.27*	*1.73*	*2.17*	*1.57*	*1.50*	**4.07**	*2.73*	**4.03**	*32,370*
Photographer	−0.11	**4.23**	3.70	**3.97**	2.57	3.40	3.23	*1.97*	3.57	**4.20**	2.80	3.83	41,340
Priest	−0.12	3.17	3.37	2.30	3.13	2.27	**4.63**	*1.77*	3.13	*3.67*	**4.27**	3.73	49,800
Volunteer	−0.15	**3.83**	*3.33*	**4.57**	*2.13*	**4.07**	**4.40**	2.07	**3.87**	*3.47*	**4.43**	**4.40**	44,310
Dancer	−0.15	3.47	3.87	3.43	*1.50*	3.57	3.17	2.73	3.10	3.77	*2.40*	2.90	41,500
Sex Worker	−0.16	2.90	*2.50*	*1.23*	*1.33*	*1.07*	*2.57*	**4.50**	3.03	3.90	*2.27*	*2.43*	**77,053**
Stockbroker	−0.16	2.83	3.87	2.50	**3.30**	3.60	*2.13*	2.07	3.03	3.87	2.80	*1.90*	59,165
CEO	−0.18	**4.20**	**4.17**	**3.60**	**4.77**	**4.43**	3.83	2.63	**4.53**	3.80	3.73	*2.17*	**397,232**
Athlete	**−0.27**	2.90	**4.40**	3.47	2.77	**4.30**	3.13	2.93	*2.47*	**4.13**	3.13	*2.30*	42,580
Doctor	**−0.31**	*2.80*	**4.17**	**3.90**	**4.40**	**4.50**	**4.60**	**3.23**	**4.13**	*3.73*	**4.90**	2.97	**74,450**
Soldier	**−0.31**	*2.27*	**3.93**	2.20	**3.70**	3.27	3.33	**4.57**	**4.10**	*3.27*	3.77	*2.27*	**83,000**
Lawyer	**−0.32**	3.20	**4.17**	3.50	**4.33**	**3.87**	**4.03**	2.20	**4.10**	3.87	**4.43**	*2.23*	**61,486**
Bouncer	**−0.33**	*2.20*	*2.57*	*1.73*	2.97	*1.73*	*2.13*	**3.50**	*2.00*	*3.60*	*2.57*	**4.23**	*37,170*
ICC		0.93	0.92	0.95	0.97	0.96	0.95	0.96	0.94	0.52	0.94	0.96	
SD		0.77	0.67	0.99	1.10	1.06	0.86	0.92	0.89	0.26	0.86	0.86	

AU, autonomy. FB, feedback. WPay, work without pay. PR, prestige. PWR, power. RISK, safety/danger. RL, relatedness. SV, skill variety. TI, task identity. TS, task significance. WLB, work life balance. Top six least semantic italicized and underlined. Top five scale values bolded.

people. Also, many holders of these jobs in Norway are people who combine this job with taking a higher education, because it often takes place outside of office hours.

Concerning the second external dataset, the panel data, we hypothesized as H3 that this dataset also would be significantly related to semantic compliance – even after controlling for salary level. We found support for this as well, but not as strongly as with the salary level. Generally, semantic compliance was visibly correlated with most of the job characteristics that also influence levels of motivation such as autonomy, feedback, power, prestige, skill variety and task significance. It is also possible to see from the distribution in **Table 6** that semantic compliance does seem related to high and low clusters along work characteristics. These effects were generally changed a bit when controlling for the salary levels, but still had visible influence on the groups' semantic compliance. Moreover, the semantic compliance of the respondents correlated significantly stronger with the panel's ratings of their jobs than with their own motivational measures. We believe this speaks strongly in favor of the semantic compliance not being a methodological artifact, even if the aggregation on group level only $n = 18$ job types raised issues of statistical significance.

Taken together, our results indicate that job characteristics and salary levels do influence self-rated levels of motivation as found in previous research, but they also influence semantic compliance independently of the score levels. The emerging differences in semantic compliance are interacting with motivational variables and job types and indicate that extensive differences in interpretation of items take place when respondents enter their scores. Job characteristics still pose the most powerful direct influence on differences in motivational levels, but the influence of semantics is sizeable and sometimes even stronger than the job types.

The theoretical and practical relevance of our findings can be seen by comparing the score levels of some of the professional groups. According to their reported score levels, CEOs are just as intrinsically motivated as priests, and claim just as little interest in their pay level. If this were true in an absolute sense, it would obviate any discussion about executive compensation, which probably is an unlikely interpretation (Ellig, 2014; Shin, 2016). Priests and sex workers differ only on 3 out of 8 measures (affective commitment, economic exchange, and IM), despite their possible differences in work values. Stockbrokers and sex workers have no score level differences but have widely different scores on job characteristics such as autonomy, relatedness, skill variety, and task identity. They work with high effort and quality, and all but bouncers, cleaners and photographers rarely think of quitting their jobs. All respondents claim to be more intrinsically motivated than interested in money (with the possible exception of cleaners).

These similarities in score levels or lack of distinct differences pose the question: Are the numerical levels really indicative of the same level of motivation? Do the measures imply invariant quantifications (Mari et al., 2017; Maul et al., 2019), or do the numbers in the responses represent endorsed statements (Drasgow et al., 2015)? Because in the latter case, responses must be treated as context-dependent interpretations.

This question opens the discussion about the nature of semantics in survey research. Words do not have fixed meanings, independent of context (Kay, 1996; Lucy, 1996; Kintsch, 2001; Sidnell and Enfield, 2012). The context of an utterance determines how it is to be understood. As outlined in the quote by Deci et al. in the introduction (Deci et al., 2017, p. 20), people with demanding and demeaning jobs who struggle to support a family and long for days away from work may interpret some items very differently from people who never worry about paying their rents. Items related to IM is probably not indifferent to this context. The reader is invited to imagine a dinner table conversation where someone says: "I work as a priest. I easily get absorbed in my work and do not think much about my income." Try to change "priest" with any other profession on the list, and most people will get a feeling that the words somehow take on different meanings.

Previous studies have shown the general semantic predictability between the motivational variables involved in this study (Arnulf et al., 2014, 2018a). A general semantic predictability among variables imply that their relationships are given *a priori* with little room to vary (Semin, 1989; Smedslund, 2002; Arnulf, 2020), such that statements about WE and quality are implicated by other statements about motivation. The obverse side of this is that once a subject chooses a value at an entry point on the scale, the values on the other scales will be given or at least restricted in variance (Feldman and Lynch, 1988; Arnulf et al., 2018b). It is striking how most respondents rate their effort and quality in the high ranges. High self-ratings of effort may be everything from true assessments via self-serving biases (Duval and Silvia, 2002), social desirability (Furnham, 1986) and unskilled unawareness (Kruger and Dunning, 1999; Ehrlinger et al., 2008; Sheldon et al., 2014). From a semantic point of view, people who agree on the scores of one variable are also expected to agree on other variables, which is what we find. In this interpretive process, the semantic influences interact with job characteristics to shape the observed scores.

There is a methodological limitation to this process, best observed in the scores of the CEOs. These people with their high incomes are a seeming exception to the rule that higher income creates higher semantic compliance, but this is probably a ceiling effect. Respondents who score very high (or very low) on all items may reduce their semantic predictability due to the restriction of statistical range. In our sample, this may be the case for photographers, CEOs, and priests. Most of these respondents tend to give such consistently high scores that differences between items are obliterated and thereby also most semantic prediction. Where all items are given similar scores, it becomes hard to detect whether the respondent read any differences into them due to restriction of range.

The most semantically predictable participants in each professional group will therefore, with very few exceptions, be the ones who score slightly lower than the others. It is only possible to be semantically predictable for respondents who vary their scores, which by necessity implies the need for some scores to be lower than others, lowering the average score levels.

Lack of semantic predictability can therefore appear due to the following three causes, with different possible remedies. First, the restriction of range in a ceiling effect where respondents

are indiscriminately enthusiastic (or disgruntled), along with any other general response set that flattens the interpretation of items. The second possibility would be a lack of verbal acuity – the respondent does not process the items properly, due to a lack of language skills or simply sloppy reading (cf. Arnulf and Larsen, 2019). In this case, the responses would contain noise. A third possibility would be systematic differences in the way items are processed (cf. Arnulf et al., 2018c), which is what we are really looking for here. Our data show signs of all three explanations.

Ceiling- or flooring effects could be avoided by better procedures in selecting items and scale options, for example by using item response theory (IRT) (van Schuur, 2017). Lack of verbal acuity could possibly be avoided by instructing respondents differently. An unpublished master thesis found that semantic compliance tended to increase when respondents were forced to delay responses with a number of seconds after having been exposed to them (Noack and Bonde, 2018). But maybe the most promising way to proceed with this line of research is to systematically assess the differences in semantic compliance the way we have begun here. Our results indicate that differences in semantic compliance is a systematic characteristic in groups, and that the impact of this is possible to assess.

Elaborating on this point, two limitations of our design are important to bear in mind. First, we are only using one single semantic space. This space seems to favor the language usage of high-status, high-income participants. The semantic algorithms here present some sort of a *standard* language usage, against which all other groups are measured. Conceivably, other groups might be predictable using other types of semantic similarity indices or from other semantic spaces. This question is treated in length by Kintsch (2001), who showed that LSA will need special procedures to pick up the usual differences in language parsing that appear in normal human speakers when contexts change. The systematic tendency for the one semantic space that we use here to predict some groups better than others is probably due to systematic differences in how contexts influence the understanding of items.

Secondly, the different professions also differ in which type of motivational scale is most likely to expose their semantic differences. The two artistic professions, artists and photographers, are usually single person businesses in our sample. Being individuals rather than organizations, the two scales commitment (AC) and organizational citizenship (OCB) create big intra-group variance because the meanings of these items may be very different or even contrived for some of them (see Schwarz, 1999). In the same vein, turnover intention (TI) may be difficult to interpret with professions such as athletes and volunteers where the subjects are probably very conscious of the fact that they are not on a lifelong career track. At the extreme end, our magazine sellers and cleaners are mostly people who probably had no initial intention to do this for a living. This could make turnover intention a complex matter for them.

Taken together, this means that semantic predictability is a group characteristic, but one that will matter more on some variables than on others. If we could establish a common ground for determining the semantic patterns of sub-groups, we could also describe the systematic differences in meaning that different groups attribute to different items.

Even if we cannot test these patterns directly for now, we are able to conclude that different groups see the items in different ways and therefore use the items differently to express their perceived motivation. When the items of a scale (or items between scales) combine to form average score levels, the classic psychometric way of treating the data is to view the numbers as indicating a composite variable. If semantics had not played a role, only scale levels would matter. In that case, the score levels could have been taken as indicators of a *dominance* model in attitude strength (Drasgow et al., 2015), because respondents would only differ along motivational levels. Semantic analyses of the items take this a step further and point to how the items are related to each other in terms of meaning. What we see in the patterns of LSA cosines is how likely one response is, given its relationship to the meaning of other responses. In our data, high-status job holders seem to share this view of the items and respond consistently. This consistent choice of responses is what Coombs called "unfolding" (Coombs and Kao, 1960), and which has been experimentally demonstrated to be highly consistent in individuals (Michell, 1994). However, when other groups of respondents display similar average score levels but deviate from the semantically expected, it means that they are sorting the response options differently. In other words, they are making different combinations of response options from the semantically expected.

This goes to the core of Likert's (1932) original problem – the relationship between stated points of view and their numerical representations. We offer respondents verbal response options ("is it very likely or very unlikely that you will look for a new job?") that we translate into numbers (1 – 5) and calculate in statistics. After arriving at the numbers, we need to interpret these into words again ("people who are mostly motivated by money are more likely to look for new jobs"). As claimed by Kjell et al. (2019), semantic algorithms may principally allow us to bypass the numbers and stay with the response texts. Looking at **Table 5**, we rounded up the mean scores to integers to represent statements about motivation. This created a picture where many job types seemed to express their motivation through fairly identical statements. This rounding up of mean scores did not only conceal significant decimal differences between the groups, it also concealed important semantic differences between the professions. The mean level of scales does not show how the mutual ranking of each item may differ between the professions – they may have ranked items differently to create different stories about their work motivation. Moreover, even similar wordings may have different meanings in different contexts. The same score on the same item seems sometimes to have a different meaning if the context differs.

Limitations

Our present design required that we varied the job types to ascertain reliable variation in the situational factors, but we restricted the variation in the survey scales that we used. All eight scales were somehow related to measuring motivation. The Cronbach's alpha of all 50 items combined is actually

0.91. With this homogeneous sample of items, the range of semantic differences is also limited. This means that the LSA cosines probably are an under-estimation of the true semantic structure of the survey. The algorithms are, at the current time, still inferior to humans in language parsing, and so the cosines will contain noise and probably miss semantic differences that are important to the human respondents. A semantically diverse survey structure would possibly make the semantic algorithms more sensitive to differences in semantics between groups. Another limitation is the sample size and the lack of cultural variation in the groups. Larger samples and samples spanning more countries than Norway might very well change the observed statistics.

CONCLUSION

We set out to examine whether the semantic response characteristics of individuals would vary across groups, and this seems to be the case. Whereas we usually would look at how different work situations or professional characteristics influence motivation, we also find that the same characteristics influence semantic parsing of item texts. Different situations produce different patterns of relating to the texts in a quantifiable way, about half as predictive of motivational levels as the job situations themselves. One may object that the motivational levels are measurements that we intend to produce – levels of motivation. The semantic patterns are not intended outcomes of the surveys and more difficult to interpret. And yet, as we have shown, the motivational levels have shortcomings seen as measurements of motivation. It is not obvious that the same numerical levels of motivation indicate the same subjective situation in different respondents. As Solomon Asch warned in his book *Social Psychology,* "most social acts have to be understood in their setting, and lose meaning if isolated. No error in thinking about social facts is more serious than the failure to see their place and function" (Asch, 1987, p. 61, orig. 1952). This also seems to apply to Likert-scale statements. The context determines the meaning of the items and influences the interpretation of score levels. Our conclusion is therefore that the semantic characteristics of individuals, the way they interpret items and take context into consideration, is a necessary and integral part of survey data.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the NSD Norsk Samfunnsvitenskapelig Datatjeneste. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JA designed the study, supervised the data collection, and co-wrote the text. KN analyzed the data, producing the tables and figures, and co-wrote the text. KL performed the semantic algorithms and co-wrote the text. CH and MA established the measures, obtained the samples, and made a preliminary analysis of the data.

REFERENCES

Allen, N. J., and Meyer, J. P. (1990). The measurement and antecedents of affective, continuance, and normative commitments to organization. *J. Occup. Psychol.* 63, 1–8.

Anandarajan, M., Hill, C., and Nolan, T. (2019). "Latent semantic analysis (LSA) in python," in *Practical Text Analytics: Maximizing the Value of Text Data* (Cham: Springer), 221–242.

Andrich, D. (1996). A hyperbolic cosine latent trait model for unfolding polytomous responses: reconciling Thurstone and Likert methodologies. *Br. J. Math. Stat. Psychol.* 49, 347–365. doi: 10.1111/j.2044-8317.1996.tb01093.x

Antonakis, J., Bendahan, S., Jacquart, P., and Lalive, R. (2010). On making causal claims: a review and recommendations. *Leadersh. Q.* 21, 1086–1120. doi: 10.1016/j.leaqua.2010.10.010

Arnulf, J. K. (2020). "Wittgenstein's revenge: How semantic algorithms can help survey research escape Smedslund's labyrinth," in *Respect for Thought; Jan Smedslund's Legacy for Psychology*, eds T. G. Lindstad, E. Stänicke, and J. Valsiner (Berlin: Springer).

Arnulf, J. K., and Larsen, K. R. (2019). Too inclusive? How Likert-scale surveys may overlook cross-cultural differences in leadership. *Paper Presented at the Academy of Management Meeting*, Boston, MA.

Arnulf, J. K., and Larsen, K. R. (2020). Culture blind leadership research: how semantically determined survey data may fail to detect cultural differences. *Front. Psychol.* 11:176. doi: 10.3389/fpsyg.2020.00176

Arnulf, J. K., Larsen, K., and Dysvik, A. (2018a). Measuring semantic components in training and motivation: a methodological introduction to the semantic theory of survey response. *Hum. Resour. Dev. Q.* 30, 17–38. doi: 10.1002/hrdq.21324

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018b). Respondent robotics: simulating responses to Likert-scale survey items. *Sage Open* 8, 1–18. doi: 10.1177/2158244018764803

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018c). Semantic algorithms can detect how media language shapes survey responses in organizational behaviour. *PLoS One* 13:e0207643. doi: 10.1371/journal.pone.0207643

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Bong, C. H. (2014). Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. *PLoS One* 9:e106361. doi: 10.1371/journal.pone.0106361

Arnulf, J. K., Larsen, K. R., Martinsen, O. L., and Egeland, T. (2018d). The failing measurement of attitudes: how semantic determinants of individual survey responses come to replace measures of attitude strength. *Behav. Res. Methods* 50, 2345–2365. doi: 10.3758/s13428-017-0999-y

Asch, S. E. (1987). *Social Psychology*. Oxford: Oxford University Press.

Bååth, R., Sikström, S., Kalnak, N., Hansson, K., and Sahlén, B. (2019). Latent semantic analysis discriminates children with developmental language disorder (DLD) from children with typical language development. *J. Psycholinguist. Res.* 48, 683–697. doi: 10.1007/s10936-018-09625-8

Barrick, M. R., Mount, M. K., and Li, N. (2013). The theory of purposeful work behavior: the role of personality, higher-order goals, and job characteristics. *Acad. Manage. Rev.* 38, 132–153. doi: 10.5465/amr.2010.0479

Benichov, J., Cox, L. C., Tun, P. A., and Wingfield, A. (2012). Word recognition within a linguistic context: effects of age, hearing acuity, verbal

ability, and cognitive function. *Ear Hear.* 33, 250–256. doi: 10.1097/AUD.0b013e31822f680f

Borsboom, D. (2008). Latent variable theory. *Measurement* 6, 25–53.

Buch, R., Kuvaas, B., and Dysvik, A. (2012). If and when social and economic leader-member exchange relationships predict follower work effort : the moderating role of work motivation. *Leadersh. Organ. Dev. J.* 35, 725–739. doi: 10.1108/lodj-09-2012-0121

Cameron, J., and Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: a meta-analysis. *Rev. Educ. Res.* 64, 363–423. doi: 10.3102/00346543064003363

Campbell, N. R. (1920). *Physics - The Elements*. Cambridge: Cambridge University Press.

Cascio, W. F. (2012). Methodological issues in international HR management research. *Int. J. Hum. Resour. Manage.* 23, 2532–2545. doi: 10.1080/09585192.2011.561242

Chiu, S. F., and Chen, H. L. (2005). Relationship between job characteristics and organizational citizenship behavior: the mediational role of job satisfaction. *Soc. Behav. Pers.* 33, 523–539. doi: 10.2224/sbp.2005.33.6.523

Coombs, C. H., and Kao, R. C. (1960). On a connection between factor-analysis and multidimensional unfolding. *Psychometrika* 25, 219–231. doi: 10.1007/Bf02289726

Cooper, L. (1969). Athletics, activity and personality: a review of the literature. *Res. Q.* 40, 17–22. doi: 10.1080/10671188.1969.10616637

Deci, E. L., Connell, J. P., and Ryan, R. M. (1989). Self-determination in a work organization. *J. Appl. Psychol.* 74, 580–590. doi: 10.1037//0021-9010.74.4.580

Deci, E. L., Olafsen, A. H., and Ryan, R. M. (2017). Self-determination theory in work organizations: the state of a science. *Annu. Rev. Organ. Psychol. Organ. Behav.* 4, 19–43.

Delecta, P. (2011). Work life balance. *Int. J. Curr. Res.* 3, 186–189.

Dennis, S., Landauer, T. K., Kintsch, W., and Quesada, J. (2013). *Introduction to Latent Semantic Analysis*. Denver, CO: University of Colorado Press.

Drasgow, F., Chernyshenko, O. S., and Stark, S. (2015). 75 years after likert: Thurstone was right! *Ind. Organ. Psychol.* 3, 465–476. doi: 10.1111/j.1754-9434.2010.01273.x

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. *Paper Presented at the Conference on Human Factors in Computing Systems*, Washington, DC.

Duval, T. S., and Silvia, P. J. (2002). Self-awareness, probability of improvement, and the self-serving bias. *J. Pers. Soc. Psychol.* 82, 49–61. doi: 10.1037/0022-3514.82.1.49

Dysvik, A., and Kuvaas, B. (2011). Intrinsic motivation as a moderator on the relationship between perceived job autonomy and work performance. *Eur. J. Work Organ. Psychol.* 20, 367–387. doi: 10.1080/13594321003590630

Dysvik, A., Kuvaas, B., and Buch, R. (2010). Trainee programme reactions and work performance: the moderating role of intrinsic motivation. *Hum. Resour. Dev. Int.* 13, 409–423. doi: 10.1080/13678868.2010.501962

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., and Kruger, J. (2008). Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent. *Organ. Behav. Hum. Decis. Process.* 105, 98–121. doi: 10.1016/j.obhdp.2007.05.002

Ellig, B. (2014). *The Complete Guide to Executive Compensation*. New York, NY: McGraw-Hill.

Elster, J. (2011). Hard and soft obscurantism in the humanities and social sciences. *Diogenes* 58, 159–170. doi: 10.1177/0392192112444984

Elvevag, B., Foltz, P. W., Rosenstein, M., Ferrer, I. C. R., De Deyne, S., Mizraji, E., et al. (2017). Thoughts about disordered thinking: measuring and quantifying the laws of order and disorder. *Schizophr. Bull.* 43, 509–513. doi: 10.1093/schbul/sbx040

Fang, M., and Gerhart, B. (2012). Does pay for performance diminish intrinsic interest? *Int. J. Hum. Resour. Manage.* 23, 1176–1196. doi: 10.1080/09585192.2011.561227

Feldman, J. M., and Lynch, J. G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *J. Appl. Psychol.* 73, 421–435. doi: 10.1037//0021-9010.73.3.421

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Pers. Individ. Differ.* 7, 385–400. doi: 10.1016/0191-8869(86)90014-0

Geertz, C. (1973). *The Interpretation of Cultures*. New York, NY: Basic Books.

Gefen, D., Endicott, J. E., Miller, J., Fresneda, J. E., and Larsen, K. R. (2017). A guide to text analysis with latent semantic analysis in r with annotated code: studying online reviews and the stack exchange community. *Commun. Assoc. Inf. Syst.* 41, 450–496. doi: 10.17705/1cais.04121

Gefen, D., and Larsen, K. (2017). Controlling for lexical closeness in survey research: a demonstration on the technology acceptance model. *J. Assoc. Inf. Syst.* 18, 727–757. doi: 10.17705/1jais.00469

Grant, A. M. (2008). Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity. *J. Appl. Psychol.* 93, 48–58. doi: 10.1037/0021-9010.93.1.48

Günther, F., Dudschig, C., and Kaup, B. (2015). LSAfun - An R package for computations based on Latent Semantic Analysis. *Behav. Res. Methods* 47, 930–944. doi: 10.3758/s13428-014-0529-0

Hackman, J. R., and Oldham, G. R. (1975). Development of the job diagnostic survey. *J. Appl. Psychol.* 60, 159–170. doi: 10.1037/h0076546

Hackman, J. R., and Oldham, G. R. (1976). Motivation through the design of work: test of a theory. *Organ. Behav. Hum. Perform.* 16, 250–279. doi: 10.1016/0030-5073(76)90016-7

Hauser, R. M., and Warren, J. R. (2012). socioeconomic indexes for occupations: a review, update, and critique. *Sociol. Methodol.* 27, 177–298. doi: 10.1111/1467-9531.271028

Kanfer, R., Frese, M., and Johnson, R. E. (2017). Motivation related to work: a century of progress. *J. Appl. Psychol.* 102, 338–355. doi: 10.1037/apl0000133

Kay, P. (1996). "Intra-speaker relativity," in *Rethinking Linguistic Relativity*, eds J. Gumperz and S. C. Levinson (Cambridge: Cambridge University Press), 97–111.

Kintsch, W. (2001). Predication. *Cogn. Sci.* 25, 173–202.

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikstrom, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Kruger, J., and Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* 77, 1121–1134. doi: 10.1037/0022-3514.77.6.1121

Kuvaas, B. (2006a). Performance appraisal satisfaction and employee outcomes: mediating and moderating roles of work motivation. *Int. J. Hum. Resour. Manage.* 17, 504–522. doi: 10.1080/09585190500521581

Kuvaas, B. (2006b). Work performance, affective commitment, and work motivation: the roles of pay administration and pay level. *J. Organ. Behav.* 27, 365–385. doi: 10.1002/job.377

Kuvaas, B., and Dysvik, A. (2009). perceived investment in permanent employee development and social and economic exchange perceptions among temporary employees. *J. Appl. Soc. Psychol.* 39, 2499–2524. doi: 10.1111/j.1559-1816.2009.00535.x

Lamiell, J. T. (2013). Statisticism in personality psychologists' use of trait constructs: What is it? How was it contracted? Is there a cure? *New Ideas Psychol.* 31, 65–71. doi: 10.1016/j.newideapsych.2011.02.009

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037//0033-295x.104.2.211

Larsen, K. R., and Bong, C. H. (2016). A tool for addressing construct identity in literature reviews and meta-analyses. *MIS Q.* 40, 529–551. doi: 10.25300/Misq/2016/40.3.01

Likert, R. (1932). A technique for the measurement of attitudes. *Arch. Psychol.* 140, 1–55.

Lucy, J. (1996). "The scope of linguistic relativity: an analysis and a review of empirical research," in *Rethinking Linguistic Relativity*, eds J. Gumperz and S. C. Levinson (Cambridge: Cambridge University Press), 37–69.

Mari, L., Maul, A., Irribarra, D. T., and Wilson, M. (2017). Quantities, quantification, and the necessary and sufficient conditions for measurement. *Measurement* 100, 115–121. doi: 10.1016/j.measurement.2016.12.050

Maul, A. (2017). Rethinking traditional methods of survey validation. *Meas. Interdiscip. Res. Perspect.* 15, 51–69. doi: 10.1080/15366367.2017.1348108

Maul, A., Mari, L., and Wilson, M. (2019). Intersubjectivity of measurement across the sciences. *Measurement* 131, 764–770. doi: 10.1016/j.measurement.2018.08.068

McClelland, D. C., Koestner, R., and Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychol. Rev.* 96, 690–702. doi: 10.1037//0033-295x.96.4.690

Meyer, J. P., and Allen, N. J. (1997). *Commitment in the Workplace: Theory, Research, and Application*. Thousand Oaks, CA: Sage Publication.

Michell, J. (1994). Measuring dimensions of belief by unidimensional unfolding. *J. Math. Psychol.* 38, 244–273. doi: 10.1006/jmps.1994.1016

Nimon, K., Lewis, M., Kane, R., and Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: an introduction to the package and a practical example. *Behav. Res. Methods* 40, 457–466. doi: 10.3758/brm.40.2.457

Nimon, K., Shuck, B., and Zigarmi, D. (2016). Construct overlap between employee engagement and job satisfaction: A function of semantic equivalence? *J. Happiness Stud.* 17, 1149–1171. doi: 10.1007/s10902-015-9636-6

Noack, E., and Bonde, C. (2018). *Leadership and Semantics: Explorations of the Semantic Theory of Survey Responses*. Master's thesis, BI Norwegian Business School, Oslo.

Obermann, J., and Velte, P. (2018). Determinants and consequences of executive compensation-related shareholder activism and say-on-pay votes: a literature review and research agenda. *J. Acc. Lit.* 40, 116–151. doi: 10.1016/j.acclit.2018.02.001

Oldham, G. R., and Hackman, J. R. (2010). Not what it was and not what it will be: the future of job design research. *J. Organ. Behav.* 31, 463–479. doi: 10.1002/job.678

Organ, D. W. (1988). *Organizational Citizenship Behavior: The Good Soldier Syndrome*. Lexington, MA: Lexington Books.

Parks-Stamm, E. J., Oettingen, G., and Gollwitzer, P. M. (2010). Making sense of one's actions in an explanatory vacuum: the interpretation of nonconscious goal striving. *J. Exp. Soc. Psychol.* 46, 531–542. doi: 10.1016/j.jesp.2010.02.004

Podsakoff, P. M., MacKenzie, S. B., and Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annu. Rev. Psychol.* 63, 539–569. doi: 10.1146/annurev-psych-120710-100452

Podsakoff, P. M., MacKenzie, S. B., Podsakoff, N. P., and Lee, J. Y. (2003). The mismeasure of man(agement) and its implications for leadership research. *Leadersh. Q.* 14, 615–656. doi: 10.1016/j.leaqua.2003.08.002

Putnick, D. L., and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev. Rev.* 41, 71–90. doi: 10.1016/j.dr.2016.06.004

Rockmann, K. W., and Ballinger, G. A. (2017). Intrinsic motivation and organizational identification among on-demand workers. *J. Appl. Psychol* .102, 1305–1316. doi: 10.1037/apl0000224

Rosenbusch, H., Wanders, F., and Pit, I. L. (2019). The Semantic Scale Network: an online tool to detect semantic overlap of psychological scales and prevent scale redundancies. *Psychol. Methods* 25, 380–392. doi: 10.1037/met0000244

Ryan, R. M., and Deci, E. L. (2000a). Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemp. Educ. Psychol.* 25, 54–67. doi: 10.1006/ceps.1999.1020

Ryan, R. M., and Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066x.55.1.68

Ryan, R. M., and Deci, E. L. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York, NY: Guilford Press.

Schwarz, N. (1999). Self-reports - How the questions shape the answers. *Am. Psychol.* 54, 93–105. doi: 10.1037//0003-066x.54.2.93

Semin, G. (1989). The contributon of linguistic factors to attribute inference and semantic similarity judgements. *Eur. J. Soc. Psychol.* 19, 85–100. doi: 10.1002/ejsp.2420190202

Sheldon, O. J., Dunning, D., and Ames, D. R. (2014). Emotionally unskilled, unaware, and uninterested in learning more: reactions to feedback about deficits in emotional intelligence. *J. Appl. Psychol.* 99, 125–137. doi: 10.1037/a0034138

Shin, T. J. (2016). Fair pay or power play? Pay equity, managerial power, and compensation adjustments for CEOs. *J. Manage.* 42, 419–448. doi: 10.1177/0149206313478186

Shore, L. M., Tetrick, L. E., Lynch, P., and Barksdale, K. (2006). Social and economic exchange: construct development and validation. *J. Appl. Soc. Psychol.* 36, 837–867. doi: 10.1111/j.0021-9029.2006.00046.x

Sidnell, J., and Enfield, N. J. (2012). Language diversity and social action a third locus of linguistic relativity. *Curr. Anthropol.* 53, 302–333. doi: 10.1086/665697

Sirota, D., Mischkind, L. A., and Meltzer, M. I. (2005). *The Enthusiastic Employee: How Companies Profit By Giving Workers What They Want*. Upper Saddle River, NJ: Pearson.

Slaney, K. L. (2017). "Some conceptual housecleaning," in *Validating Psychological Constructs: Historical, Philosophical, and Practical Dimensions* (London: Palgrave Macmillan), 201–234. doi: 10.1057/978-1-137-38523-9_8

Smedslund, J. (1988). What is measured by a psychological measure. *Scand. J. Psychol.* 29, 148–151. doi: 10.1111/j.1467-9450.1988.tb00785.x

Smedslund, J. (2002). From hypothesis-testing psychology to procedure-testing psychologic. *Rev. Gen. Psychol.* 6, 51–72. doi: 10.1037/1089-2680.6.1.51

Van Dyne, L., and LePine, J. A. (1998). Helping and voice extra-role behaviors: evidence of construct and predictive validity. *Acad. Manage. J.* 41, 108–119. doi: 10.5465/256902

van Schuur, W. H. (2017). Mokken scale analysis: between the Guttman scale and parametric item response theory. *Polit. Anal.* 11, 139–163. doi: 10.1093/pan/mpg002

Weick, K. E. (1995). *Sensemaking in Organizations*. Thousand Oaks, CA: Sage.

Weick, K. E. (2012). Organized sensemaking: a commentary on processes of interpretive work. *Hum. Relat.* 65, 141–153. doi: 10.1177/0018726711424235

Wild, F. (2015). *Package 'lsa': CRAN*. Available online at: https://cran.r-project.org/web/packages/lsa/lsa.pdf (accessed May 1, 2020).

Yukl, G. (2012a). Effective leadership behavior: what we know and what questions need more attention. *Acad. Manage. Perspect.* 26, 66–85. doi: 10.5465/amp.2012.0088

Yukl, G. (2012b). *Leadership in Organizations*, 8th Edn. Harlow: Pearson Education.

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

frontiers in Psychology

DATA REPORT
published: 03 February 2021
doi: 10.3389/fpsyg.2020.523494

MOWDOC: A Dataset of Documents From *Taking the Measure of Work* for Building a Latent Semantic Analysis Space

*Kim F. Nimon**

Human Resource Development, University of Texas at Tyler, Tyler, TX, United States

Keywords: latent semantic analysis, semantic survey response theory, surveys, jingle-jangle fallacies, work, organization

INTRODUCTION

For organizational researchers employing surveys, understanding the semantic link between and among survey items and responses is key. Researchers like Schwarz (1999) have long understood, for example, that item order can impact survey responses. To account for "item wording similarity," researchers may allow item error variances to correlate (cf. Rich et al., 2010, p. 625). Other researchers, such as Newman et al. (2010), have pointed to semantic similarity between items as support for the premise that work engagement is like old wine in a new bottle.

Recently, organizational researchers (e.g., Arnulf et al., 2014, 2018) have been able to use latent semantic analysis (LSA) and semantic survey response theory (SSRT) to quantify the semantic similarity between and among scales, items, and survey responses. Latent semantic analysis is a computational model that assesses similarity in language where the similarity of any "given word (or series of words) is given by the context where this word is usually found" (Arnulf et al., 2020, p. 4). Latent semantic analysis involves establishing a semantic space from a corpus of existing documents (e.g., journal articles, newspaper stories, item sets). The corpus of documents is represented in a word-by-document matrix and then transformed into an LSA space through singular value decomposition. The reduced LSA space can be used to assess the semantic similarity of documents within the space as well as new documents that are projected onto the space.

Patterns of semantic similarity resulting from LSA have accounted for a substantive amount of variability in how individuals respond to survey items that purport to measure (a) transformational leadership, motivation, and self-reported work outcomes (60–86%; Arnulf et al., 2014), (b) employee engagement and job satisfaction (25–69%; Nimon et al., 2016), and (c) perceptions of a trainee program, intrinsic motivation, and work outcomes (31–55%, Arnulf et al., 2019). It also appears that personality, demographics, professional training, and interest in the subject matter have an impact on the degree to which an individual's responses follow a semantically predictable pattern (Arnulf et al., 2018; Arnulf and Larsen, 2020, Arnulf et al., 2020). While being able to objectively access the degree to which survey responses are impacted by semantics is a great step forward in survey research, such research is often conducted with LSA spaces that are not open and therefore not customizable except by those that have access to the body of text upon which the LSA space is built. In this day of open science, researchers need access not only to the LSA space on which semantic survey research may be based but also to the underlying corpus of text to determine whether choices made in the generation of the LSA space have an impact on the results found.

Researchers may not be able to create their own LSA spaces for a number of reasons, including the fact that on some occasions it is difficult to collect a representative corpus of text (Quesada, 2011). However, building an LSA space allows researchers to customize the space including the

OPEN ACCESS

Edited by:
Holmes Finch,
Ball State University, United States

Reviewed by:
Laszlo Hunyadi,
University of Debrecen, Hungary
Fritz Günther,
University of Milano-Bicocca, Italy

***Correspondence:**
Kim F. Nimon
kim.nimon@gmail.com

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

Received: *29 December 2019*
Accepted: *15 December 2020*
Published: *03 February 2021*

Citation:
Nimon KF (2021) MOWDOC: A Dataset of Documents From Taking the Measure of Work for Building a Latent Semantic Analysis Space. Front. Psychol. 11:523494. doi: 10.3389/fpsyg.2020.523494

application of weighting schemes and the level of dimensionality for the LSA space. As shown by Arnulf et al. (2018), the dimensionality of the LSA space is a factor when using an LSA space to predict empirical correlations from scale item cosines. To help address the barrier to creating an LSA space for use in the analysis of scale items in organizational research, this report provides a dataset of documents from measures reviewed in *Taking the Measure of Work*. In *Taking the Measure of Work*, Fields provided the items for 324 scales and subscales which cover the areas of job satisfaction, organizational commitment, job characteristics, job stress, job roles, organizational justice, work-family conflict, person-organization fit, work behaviors, and work values. The MOWDOC dataset presented in this manuscript provides the documents necessary to create a semantic space from the item sets presented in Fields's *Taking the Measure of Work*.

MOWDOC

The dataset presented in this manuscript can be accessed via https://doi.org/10.6084/m9.figshare.13298165. The dataset contains five variables for each of the 324 scales and subscales in Fields (2002). The variable *ScaleName* identifies the name of the measure as reported in Fields as well as subscale(s) as appropriate, where subscale names are preceded by a colon. The variable *ScaleRef* identifies the reference from which Fields obtained the items.

The variable *ScaleID* is a unique identifier for each scale/subscale. The first two characters of *ScaleID* identify the type of measure as delineated by Fields (2002), where JS denotes job satisfaction, OC organizational commitment, JC job characteristics, JT job stress, JR job roles, OJ organizational justice, WC work-family conflict, PO person organization fit, WB work behaviors, and WV work values. The next three characters identify the page number on which the item set first appears in Fields. The remaining characters denote subscale(s) as appropriate.

The variable *ScaleDoc* contains the document text for each scale/subscale. The scale documents were created as follows. Item texts and associated metadata from Fields (2002) were manually entered into a comma delimited file and verified by an independent and separate individual. To create the variable *ScaleDoc*, an R script was used to create a character vector by merging all item texts for a given scale/subscale where measures containing multiple item sets or subscales were treated as separate documents. The character vector was tokenized using the *tokens* function from the *quanteda* package (Benoit et al., 2018), which also removed all characters in the Unicode "Punctuation" [P] class. The tokens were then sorted so as to not violate the copyright of the scale publishers. Finally, the tokens were merged into a single character vector.

The variable *ScaleSize* identifies the number of words for each measure that ranges from 3 to 563. The *hedonism* subscale from the Work Values Survey (Schwartz, 1994) has the fewest with two items and the Inventory of Stressful Events (Motowidlo et al., 1986) has the largest with 45 items. The mean number of words across all scales is 67 with an SD of 60. Across all 324 documents, there are a total of 21,741 words.

EXAMPLE USAGE

The R code that demonstrates how the MOWDOC dataset can be used to create an LSA space and fold a new scale[1] into the created LSA space can be accessed at https://doi.org/10.6084/m9.figshare.13298177. In general, the code follows the example in Wild (2007) and the Wild (2015) demonstration of the famous Landauer et al. (1998) example. Document-feature matrices were created using the *dfm* function from the *quanteda* package (Benoit et al., 2018), rather than using the *textmatrix* function in the *lsa* package (Wild, 2015). Amongst other differences, the *dfm* function optimally creates a sparse matrix of documents and features.

Here is the R code following a typical LSA process:

First, a text matrix was constructed using the input text. In the demonstration provided, five different document-feature matrices and associated word clouds were created to illustrate the nuances associated with stemming words and removing stop words.

Second, an LSA space with full dimensionality was created and used to verify that the document-feature matrix could be reproduced.

Third, an LSA space with reduced dimensionality was created.

Fourth, document-to-document correlations and cosines were computed using the original document-feature matrix and the reduced LSA space.

Fifth, a new document was folded into the reduced LSA space.

Sixth, correlations and cosines with the new document were created.

STRENGTHS AND LIMITATIONS

The MOWDOC datasets contains the item texts for the scales/subscales in the book of *Taking the Measure of Work*. With this dataset, researchers can customize their LSA spaces to fit their research interests including the consideration of stop words, word stemming, and weighting schemes. Note, for example, the differences in the word clouds represented in **Figure 1** that result when the MOWDOC dataset was used to generate a document-feature matrix with different parameters. Not only did each document-feature matrix contain a different number of features[2], the word most frequently used across multiple scales

[1]The file JS.csv contains the items for the Hackman and Oldham (1980) job satisfaction scale and can be accessed at https://doi.org/10.6084/m9.figshare.13298168

[2]The document-feature matrix with no stemming or removal of stop words yielded 2,564 features (7.9 features on average per scale) and was 98.5% sparse. The document-feature matrix with no stemming and English stop words from the *lsa* package (Wild, 2015) removed yielded 2,253 features (7.0 features on average per scale) and was 99.2% sparse. The document-term matrix with no stemming and English stop words from the *quanteda* package (Benoit et al., 2018) removed yielded 2,433 features (7.5 features on average per scale) and was 99.0% sparse. The document-term matrix with stemming and English stop words from the *quanteda* package (Benoit et al., 2018) removed yielded 1,704 features (5.3 features on average per scale) and was 98.6% sparse.

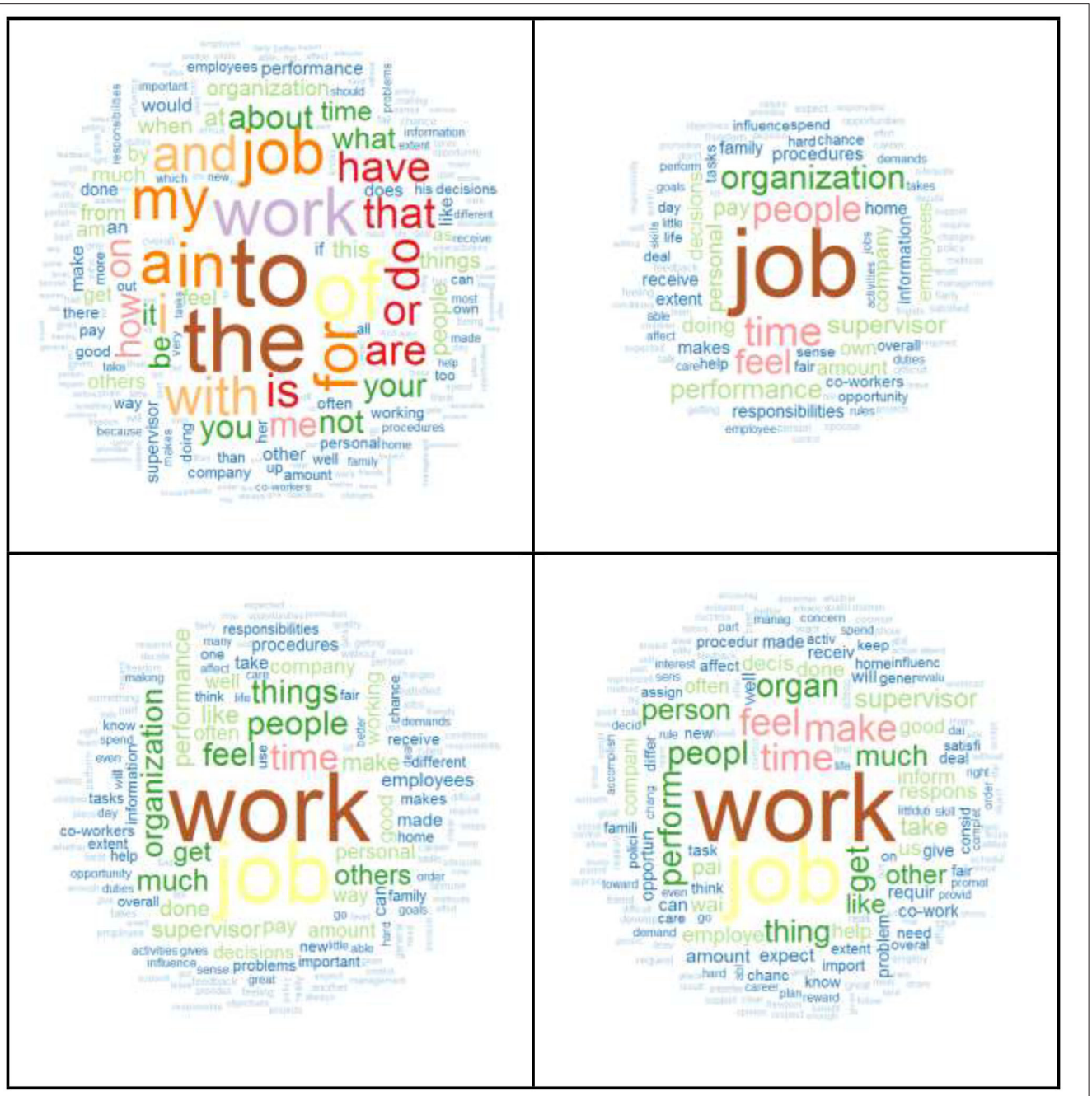


FIGURE 1 | Word clouds on MOWDOC document-feature matrices. Upper-left figure based on matrix with no stemming or stop words removed. Upper-right figure based on matrix with no stemming and English stop words from the *lsa* package (Wild, 2015) removed. Lower-left figure based on matrix with no stemming and English stop words from the *quanteda* package (Benoit et al., 2018) removed. Lower-right figure based on matrix with stemming and English stop words from the *quanteda* package (Benoit et al., 2018) removed.

differed according to how the document texts were "sanitized" (cf. Wild, 2007). In the matrix that hasn't been sanitized with no stemming or removal of stop words, "to" occurs in 76% (247) of the scales/subscales. When the matrix has no stemming but English stop words from the *lsa* package (Wild, 2015) are removed, "job" occurred in 48% (157) of the scales/subscales. A matrix with no stemming although English stop words from the *quanteda* package (Benoit et al., 2018) have been removed, "work" occurred in 53% (172) of the scales/subscales. With stemming and English stop words from the *quanteda* package (Benoit et al., 2018) removed, "work" occurred in 56% (181) of the scales/subscales. While it should not come as a surprise that

"work" is the predominant word used across scales published in a book that considers the "Measure of Work," it could be considered problematic to create an LSA space where such a relevant word was removed.

While making the document texts upon which to build an LSA space available is a strength, it might also be a limitation as resulting LSA spaces may yield over-fitted solutions when researchers assess the semantic similarity of item sets (cf. Larsen et al., 2008). It might also be a limitation that the document-feature matrices from the MOWDOC dataset tend to be sparse. Across the different "sanitization" schemes previously outlined, all matrices were at least 98.4% sparse. The dataset is also limited in that it did not preserve the word order of the original item sets. As a reviewer noted, this limits the use of the dataset to document-based models like LSA. In addition, the dataset is small for a source corpus for LSA. With 324 documents and 2,564 unique words, the use of the MOWDOC dataset may be limited beyond the example usage presented.

Clearly more research is needed to determine how the MOWDOC dataset can validly be used to inform survey research. However, even with the stated limitations, the MOWDOC dataset appears to be useful. Take for example the lsaCos.csv file that results from running the demonstration code located at https://doi.org/10.6084/m9.figshare.13298177. It yields the cosines between scales/subscales from the LSA space that was built using a document-feature matrix that was stemmed and void of English stop words contained in the *quanteda* package (Benoit et al., 2018). Notably, the cosine between the *OCBO* item set Williams and Anderson (1991, WB241B) and the *generalized compliance* item set from Smith et al. (1983, WB245B) is 0.92. Interestingly, the cosine reflects the fact that some of the items representing *OCBO*, including "attendance at work is above the norm" and "great deal of time spent with personal phone conversation," were selected from the Smith et al. (1983) generalized compliance scale.

Researchers might also fold additional items sets onto the LSA space built from *Taking the Measure of Work* to assess their semantic similarity with item sets presented in Fields (2002). For example, folding the Hackman and Oldham (1980) *job satisfaction* item set into the LSA space previously described yields a high cosine (0.86) with the *general satisfaction* item set from Jackman and Oldham (1974). Future work could include adding item texts from other compendiums of organizational research scales including those of Cook et al. (1981), Price and Mueller (1986), and Hersen and Thomas (2003), as well as submitting the existing dataset to the Semantic Scale Network offered by Rosenbusch et al. (2020).

AUTHOR CONTRIBUTIONS

The author confirms being the sole contribution of this work and has approved it for publication.

REFERENCES

Arnulf, J. K., Dysvik, A., and Larsen, K. (2019). Measuring semantic components in training and motivation: a methodological introduction to the semantic theory of survey response. *Hum. Resour. Develop. Q.* 30, 17–38. doi: 10.1002/hrdq.21324

Arnulf, J. K., and Larsen, K. R. (2020). Culture blind leadership research: how semantically determined survey data may fail to detect cultural differences. *Front. Psychol.* 11:176. doi: 10.3389/fpsyg.2020.00176

Arnulf, J. K., Larsen, K. R., and Martinsen, Ø. L. (2018). Semantic algorithms can detect how media language shapes survey responses in organizational behaviour. *PLoS ONE* 13:e0207643. doi: 10.10371/journal.pone.0207643

Arnulf, J. K., Larsen, K. R., Martinsen, Ø. L., and Bong, C. H. (2014). Predicting survey responses: how and why semantics shape survey statistics on organizational behaviour. *PLoS ONE* 9:e106361. doi: 10.10371/journal.pone.0106361

Arnulf, J. K., Nimon, K., Larsen, K. R., Hovland, C. V., and Arnesen, M. (2020). The priest, the sex worker, and the CEO: measuring motivation by job type. *Front. Psychol.* 11:1321. doi: 10.3389/fpsyg.2020.01321

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., et al. (2018). quanteda: An R package for the quantitative analysis of textual data. *J. Open Sour. Softw.* 3:774. doi: 10.21105/joss.00774

Cook, J. D., Hepworth, S. J., Wall, T. D., and Warr, P. B. (1981). *The Experience of Work: A Compendium of 249 Measures and Their Use*. London: Academic Press.

Fields, D. L. (2002). *Taking the Measure of Work: A Guide to Validated Scales for Organizational Research and Diagnosis*. Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781452231143

Hackman, J. R., and Oldham, G. R. (1980). *Work Redesign*. Reading: Addison-Wesley.

Hersen, M., and Thomas, J. C. (Eds.). (2003). *Comprehensive Handbook of Psychological Assessment, Volume 4: Industrial/Organizational Assessment*. New York, NY: Wiley.

Jackman, J. R., and Oldham, G. R. (1974). The job diagnostic survey: an instrument for the diagnosis of jobs and the evaluation of job redesign project. Tech. Rep. No. 4. New Haven, CT: Yale University, Department of Administrative Sciences.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Proces.* 25, 259–284. doi: 10.1080/016385380954028

Larsen, K., Nevo, D., and Rich, E. (2008). "Exploring the semantic validity of questionnaire scales," in *Proceedings of the 41st annual Hawaii International Conference on System Sciences* (Waikoloa, HI), 1–10.

Motowidlo, S. J., Packard, J. S., and Manning, M. R. (1986). Occupational stress: its causes and consequences for job performance. *J. Appl. Psychol.* 71, 618–629. doi: 10.1037/0021-9010.71.4.618

Newman, D. A., Joseph, D. L., and Hulin, C. L. (2010). "Job attitudes and employee engagement: considering the attitude "Afactor," in *The Handbook of Employee Engagement: Perspectives, Issues, Research, and Practice,* eds S. L. Albrecht (Northampton, MA: Edward Elgar), 43–61.

Nimon, K., Shuck, B., and Zigarmi, D. (2016). Construct overlap between employee engagement and job satisfaction: a function of semantic equivalence? *J. Happiness Stud.* 17, 1149–1171. doi: 10.1007/s10902-015-9636-6

Price, J., and Mueller, C. (1986). *Handbook of Organizational Measurement*. Marshfield, MA: Pittman.

Quesada, J. (2011). "Creating your own LSA spaces," in *Handbook of Latent Semantic Analysis,* T. K. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch (New York, NY: Routledge), 71–88.

Rich, B. L., Lepine, J. A., and Crawford, E. R. (2010). Job engagement: antecedents and effects on job performance. *Acad. Manag. J.* 53, 617–635. doi: 10.5465/amj.2010.51468988

Rosenbusch, H., Wanders, F., and Pit, I. L. (2020). The semantic scale network: an online tool to detect semantic overlap of psychological scales and

prevent scale redundancies. *Psychol. Methods* 25, 380–392. doi: 10.1037/met0000244

Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *J. Soc. Issues* 50, 19–45. doi: 10.1111/j.1540-4650.1994.tb01196.x

Schwarz, N. (1999). Self-reports - how the questions shape the answers. *Am. Psychol.* 54, 93–105. doi: 10.1037/0003-066X.54.2.93

Smith, C. A., Organ, D. W., and Near, J. P. (1983). Organizational citizenship behavior: its nature and antecedents. *J. Appl. Psychol.* 68, 653–663. doi: 10.1037/0021-9010.68.4.653

Wild, F. (2007). "An LSA package for R," in *Proceedings of the 1st International Conference on Latent Semantic Analysis in Technology Enhanced Learning (LSA-TEL'07)* (Heerlen), 11–12.

Wild, F. (2015). *lsa: Latent Semantic Analysis* (R package version 0.73.1) [Computer software]. Retrieved from: https://CRAN.R-project.org/package=lsa (accessed November 28, 2020).

Williams, L. J., and Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *J. Manag.* 17, 601–617. doi: 10.1177/014920639101700305

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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 11 May 2021
doi: 10.3389/fpsyg.2021.601679

Computational Language Assessments of Harmony in Life — Not Satisfaction With Life or Rating Scales — Correlate With Cooperative Behaviors

Oscar Kjell, Daiva Daukantaitė† and Sverker Sikström*

Department of Psychology, Lund University, Lund, Sweden

Different types of well-being are likely to be associated with different kinds of behaviors. The first objective of this study was, from a subjective well-being perspective, to examine whether harmony in life and satisfaction with life are related differently to cooperative behaviors depending on individuals' social value orientation. The second objective was, from a methodological perspective, to examine whether language-based assessments called *computational language assessments* (CLA), which enable respondents to answer with words that are analyzed using natural language processing, demonstrate stronger correlations with cooperation than traditional rating scales. Participants reported their harmony in life, satisfaction with life, and social value orientation before taking part in an online cooperative task. The results show that the CLA of overall harmony in life correlated with cooperation (all participants: $r = 0.18$, $p < 0.05$, $n = 181$) and that this was particularly true for prosocial participants ($r = 0.35$, $p < 0.001$, $n = 96$), whereas rating scales were not correlated ($p > 0.05$). No significant correlations (measured by the CLA or traditional rating scales) were found between satisfaction with life and cooperation. In conclusion, our study reveals an important behavioral difference between different types of subjective well-being. To our knowledge, this is the first study supporting the validity of self-reported CLA over traditional rating scales in relation to actual behaviors.

Keywords: natural language processing (NLP), cooperation, satisfaction with life, computational language assessments, harmony in life

OPEN ACCESS

Edited by:
Kai R. Larsen,
University of Colorado Boulder,
United States

Reviewed by:
Lusilda Schutte,
North-West University, South Africa
Admassu Nadew Lamu,
University of Bergen, Norway

****Correspondence:***
Oscar Kjell
oscar.kjell@psy.lu.se

†ORCID:
Daiva Daukantaitė
orcid.org/0000-0002-1994-041X

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

Received: *01 September 2020*
Accepted: *31 March 2021*
Published: *11 May 2021*

Citation:
Kjell O, Daukantaitė D and Sikström S (2021) Computational Language Assessments of Harmony in Life — Not Satisfaction With Life or Rating Scales — Correlate With Cooperative Behaviors. Front. Psychol. 12:601679. doi: 10.3389/fpsyg.2021.601679

INTRODUCTION

Different types of well-being are proposed to be associated with different kinds of behaviors (e.g., Ryan and Deci, 2001; Kjell, 2011). Individuals associate the pursuit of harmony in life with *cooperation* and related words (e.g., *together*, *unity*, and *mutual*), whereas the pursuit of satisfaction with life is associated with words relating to self-fulfilment (e.g., *achievement*, *goals*, and *winning*; Kjell et al., 2016). This distinction is also found when having participants describe their *level* (rather than *pursuit*) of harmony in life versus satisfaction with life using open-ended language-based measures, but not when using traditional numeric rating scales (Kjell et al., 2019). The present study allowed individuals to describe their well-being in their own words and had two objectives. The first objective was related to *well-being and cooperation*, i.e., to examine if two cognitive

components of subjective well-being – namely, overall harmony in life and overall satisfaction with life – as reported prior to a social dilemma game are related to cooperative behaviors [while also controlling for values in the form of their social value orientation (SVO)]. The second objective was related to the assessment *method*, i.e., to examine whether quantitative open-ended language-based assessments (referred to as *computational language assessments*) more clearly than rating scales distinguish between harmony in life and satisfaction with life in relation to the behavioral outcome of cooperation. These objectives were examined in a social dilemma game where cooperating increased the joint outcome and not cooperating gave one the chance to personally achieve the highest outcome.

Satisfaction, Harmony, and Cooperation

The definitions of satisfaction with life and harmony in life as well as related empirical research suggest that harmony in life is more related to cooperative behaviors than satisfaction with life. Diener et al. (1985) highlight that satisfaction with life concerns a "cognitive, judgmental process" (p. 71) regarding a person's evaluation of their life situation as a whole. As such, satisfaction with life is defined as having surroundings and circumstances according to one's expectations and ideals and in accordance with one's own criteria (Diener et al., 1985). Harmony in life, on the other hand, relates to being in balance and fitting in with one's surroundings and circumstances (e.g., see Kjell et al., 2016; Kjell and Diener, 2020). Li (2006) stresses that harmony entails favorable relationships, and Li (2008) points out that "harmony is by its very nature relational. It is through mutual support and mutual dependence that things flourish" (p. 427). Considering the different definitions, harmony in life and satisfaction with life are likely to be associated with different actions and behaviors (e.g., see Kjell, 2011).

Indeed, empirical research demonstrates differences in how individuals view their pursuit of harmony in life and satisfaction with life. In a direct comparison between harmony in life and satisfaction with life, Kjell et al. (2016) revealed that participants describe their pursuit of harmony in life with words relating to interconnectedness with other people (e.g., *peace*, *balance*, *cooperation*, *unity*, *agreement*, *accord*, *concord*, *together*, *friendship*, and *forgiveness*). Meanwhile, the pursuit of satisfaction with life is described with words relating to self-centered (cf. one's own criteria) mastery (e.g., *money*, *achievement*, *wealth*, *gratification*, *goals*, *work*, *career*, *winning*, *success*, and *job*). Similarly, Kjell et al. (2019) demonstrate that many of these aspects can also be seen in participants' descriptions of their personal state of harmony in life versus satisfaction with life.

Cooperation in the Give-Some Dilemma Game

Degree of cooperation in this study was measured in a one-shot give-some dilemma game (GSDG; e.g., see Van Lange and Kuhlman, 1994). In this dilemma game participants are given an amount of money and grouped into pairs. In a simultaneous interaction, they choose to keep their money or give some or all of it to the other person. Participants are informed that any money that is given away doubles in value for the receiver. Hence, keeping the money increases the chance to personally get the highest amount (cf. satisfaction with life), while giving the money to the other person increases the joint outcome (cf. harmony in life). Participants are informed about the other's decision at the same time. Degree of cooperation is thus operationalised as the amount of money each participant decides to give away. The amount of money participants give is hypothesized to be related to their reported level of harmony in life and satisfaction with life in addition to other factors such as their SVO as discussed next.

Prosocials and Proselfs

An individual's SVO is a stable characteristic that predicts the degree of cooperation in social dilemmas (Van Lange and Kuhlman, 1994; VanLange et al., 1997; Balliet et al., 2009), and it is defined as an individual's preference for a specific resource allocation between others and oneself (McClintock, 1972). Even though individuals can be categorized into several different SVOs, at least three are typically identified: (1) individuals with a *cooperative* SVO who focus on maximizing the joint outcome for self and others, (2) individuals with a *competitive* SVO who focus on maximizing their own outcome relative to others, and (3) individuals with an *individualistic* SVO who focus on maximizing their own outcome with little or no consideration of the outcome for others (VanLange et al., 1997). Individuals categorized with a cooperative SVO are often referred to as *prosocials*, while individuals categorized with an individualistic or competitive SVO are referred to as *proselfs*.

Because SVOs are found to be stable motivations, this distinction has played an important role in research investigating various situational and contextual variables in relation to cooperation. For example, it was demonstrated that inducing guilt (as compared with a neutral state) in participants categorized as proselfs increases cooperation in a prisoner's dilemma game (Ketelaar and Au, 2003) and a GSDG (de Hooge et al., 2007). In another study using a GSDG, it was demonstrated that inducing fear (as compared with a neutral state) decreases cooperation in prosocials (Nelissen et al., 2007). Further, Kjell and Thompson (2013) compared joy, guilt, and a neutral condition within a prisoner's dilemma game. That study revealed a significant relationship between cooperation and SVO, but no significant differences in regard to the emotional conditions. It was suggested that cognitive resources and strategies (cf. the cognitive subjective well-being components of harmony in life and satisfaction with life) rather than experimentally induced emotions may have a stronger influence on cooperation. To our knowledge, there are no studies comparing the effect of the cognitive components of subjective well-being (i.e., harmony in life versus satisfaction with life) and their respective relationship to cooperation.

Open-Ended Computational Language Assessments Versus Numerical Rating Scales

Subjective well-being is typically measured using scales comprising items (e.g., *I am satisfied with life*; Diener et al., 1985)

with a closed-ended response format (e.g., ranging from *1 = Strongly disagree* to *7 = Strongly agree*). In contrast, Kjell et al. (2019) developed computational language assessments that allow respondents to answer questions about psychological constructs with words that are analyzed using natural language processing. This method enables both *measuring* as well as *describing* the psychological construct under investigation. Importantly it was shown that computational language assessments, as compared to traditional numerical rating scales, discriminate more clearly between harmony in life and satisfaction with life. For example, the numerical rating scales Harmony in Life Scale (HILS; Kjell et al., 2016) and Satisfaction With Life Scale (SWLS; Diener et al., 1985) were strongly correlated, whereas the computational language assessments of harmony in life and satisfaction with life were only moderately correlated. Furthermore, plotting the word responses demonstrated clear differences between words relating to harmony versus satisfaction when plotting according to the computational language assessments. That is, covarying the computational language assessments of harmony in life with satisfaction with life (or vice versa) when plotting significant words yielded a clear independence between the constructs. Interestingly, these differences between harmony and satisfaction were not clear when discriminating between the words using numerical rating scales rather than semantic similarity scales, nor were they clear when covarying the corresponding numerical rating scales. This discriminative property of computational language assessments suggests that they more clearly than numerical rating scales can predict behavioral outcomes that are relevant for one, but not another, psychological construct such as harmony in life and satisfaction with life.

Objectives and Hypotheses

The study had two objectives. The first objective was to examine if overall harmony in life and overall satisfaction with life reported before a social dilemma game are related to cooperative behaviors (while also controlling for values in the form of their SVO). The hypotheses related to this objective concerned how differently the pre-interaction language-based and numerical measures of harmony in life and satisfaction with life are related to cooperation in the GSDG depending on the individual's SVO.

> H_1. Level of overall harmony in life correlates positively with cooperation, especially in those categorized as prosocial.
>
> H_2. Level of overall satisfaction with life correlates negatively with cooperation, especially in those categorized as proself.

The second objective was to examine whether *computational language assessments*, as compared with rating scales, more clearly distinguish between harmony in life and satisfaction with life in regard to cooperation in the GSDG. This is, for example, based on evidence showing that computational language assessments, as compared with numerical rating scales, discriminate more clearly between constructs (Kjell et al., 2019). Therefore, it was hypothesized that:

> H_3. Computational language assessments discern the predictions in H_1 and H_2 more strongly than numerical rating scales (i.e., they reveal stronger correlations).
>
> H_4. The relationships in H_1 and H_2 are also discerned using keyword plots based on the computational language assessments. Descriptive words that participants use to describe their overall harmony in life (e.g., *peaceful* and *balance*) are associated with high cooperation, whereas words describing their overall satisfaction with life (e.g., *happy* and *fulfilled*) are associated with low cooperation.

MATERIALS AND METHODS

Participants

Participants were recruited from Amazon's Mechanical Turk, a website that enables one to pay participants to partake in studies (Paolacci et al., 2010; Mason and Suri, 2012). A total of 200 participants were recruited at once, before starting the analyses. The size of the sample was based on an 80% power to detect a correlation of $r = 0.2$ (alpha level = 0.05, two-sided), which is a correlational size that can be considered theoretically relevant for the investigated hypothesized positive correlation between harmony in life and cooperation. Four participants were removed due to failing to correctly respond to control items (a method that has been shown to increase the statistical power and reliability of datasets; e.g., see Oppenheimer et al., 2009), two were removed for raising suspicion of responding insincerely and not answering the questions independently[1], and 13 were removed for being suspicious about the authenticity of the interaction in the last feedback question (see section "Material"). The final sample consisted of 181 participants (female = 86; male = 94; other = 1) with a mean age of 34.34 ($SD = 10.21$; range = 19–63) years and a mean of 4.6 ($SD = 1.7$) on the perceived financial situation scale (range 1 = "Our income does not cover our needs, there are great difficulties" to 7 = "Our income covers our needs, and we can save"). Participants mainly came from the United States (United States = 156; India = 20; other countries = 5).

[1]The two participants were removed for raising suspicion of not completing the study independently. The two participants answered all the word-response questions identically – three of the four questions they answered by repeatedly writing "yes," and in response to the overall harmony in life question both had, in the same order, answered: "*good*, *marvelous*, *kudos*, *extraordinary*, *elegant*, *resplendent*, *enormous*, *glory*, *stupendous*, *sumptuous*." On the HILS and the SWLS, one reported a total score of 35 on both scales, and the other a score of 33 on both scales. Further, both reported the same on the demographic questions, gave $1 in the interaction, and answered the third presented alternative on the Triple-Dominance Measure (hence the potential insincerity). They reported different worker IDs but had very similar, overlapping start and submit times. Although keeping these participants in the study lowered the overall correlation between the computational language assessment of overall harmony in life and cooperation, it did not considerably affect the other correlations. Further it did not change the remaining parts of the results as the participants' SVOs were uncategorized (i.e., not categorized as prosocial or proself).

Material

Rating Scales Measures

The HILS (Kjell et al., 2016) consists of five items (e.g., "I am in harmony") answered on a 7-point scale ranging from 1 = Strongly disagree to 7 = Strongly agree. Cronbach's alpha in the current study was 0.94 (Mc Donald's ω total = 0.96).

The SWLS (Diener et al., 1985) comprises five items (e.g., "I am satisfied with my life") answered on the same scale as the HILS. Cronbach's alpha in the current study was 0.92 (McDonald's ω total = 0.95).

The *Triple-Dominance Measure* (TDM; VanLange et al., 1997) was used to assess SVO. The TDM comprises nine items, which each present three distributions of "valuable points" that are differently shared between the respondent and a hypothetical unknown other person. Distributions with equal division of valuable points are categorized as *prosocial*, and distributions where respondents get more than the other are categorized as *proself*. If six or more answers consistently fall within one of the categories, the respondent are classified accordingly.

Demographic questions included gender, age, first language, and country of origin as well as perceived financial situation (i.e., "Does the total income of your household allow you to cover your needs?"; answered on a scale ranging from 1 = "Our income does not cover our needs, there are great difficulties" to 7 = "Our income covers our needs, and we can save").

The *control items* "On this question please answer the alternative 'neither agree nor disagree"' and "Answer 'disagree' on this question" were included with the numerical rating scales to ensure that the participants had read the questions within the survey. Participants that did not answer these items correctly were removed from the analyses. This kind of method has been demonstrated to ensure high statistical power and reliability (e.g., see Oppenheimer et al., 2009).

Word and Text Measures

The *Word-Response Harmony Question* (Kjell et al., 2019) is stated as "Overall in your life, are you in harmony or not?" The *Word-Response Satisfaction Question* (Kjell et al., 2019) reads "Overall in your life, are you satisfied or not?" These word-response questions are presented with the instructions to answer using 10 descriptive words for each question (for full instructions, see Kjell et al., 2019).

A Feedback Question asked participants to provide a brief description of their thoughts regarding the GSDG. Three psychology researchers (two with a Ph.D. and one Ph.D. student) not involved in the study, and blind to how the participants responded to other questions, evaluated the answers based on whether they raised any suspicion that the interaction did not involve another person. Participants were removed when at least two out of the three assessors indicated raised suspicion. In total 13 participants were removed (all three assessors agreed on 12 answers and on 1 answer two raters indicated suspicion; only one other answer was indicated as raising suspicion by one assessor, which was thus kept).

The *Affective Norms for English Words* (ANEW; Bradley and Lang, 1999) enabled the construction of language predicted valence scales (see the section on "Natural Language Processing and Statistical Analyses"). These affective norms comprise a large number of words that have been rated by individuals in terms of valence, arousal, and dominance. The valence model used in this study to predict the valence of responses demonstrated a cross-validated Pearson r of 0.73 ($p < 0.001$, $N = 1025$).

Intervention

The GSDG (Van Lange and Kuhlman, 1994; de Hooge et al., 2007) involved giving each participant \$1.0 and the option to give the money to an interacting partner who simultaneously had the same opportunity. However, the experiment involved a deception in which the "partner" consisted of a computer that randomly responded by either giving \$0 or \$1. Participants were informed that the amount they decided to give away would double in value for the receiver but that none of the parties in the interaction would know in advance what the other decided to give. The available alternatives to give ranged from \$0 to \$1, with \$0.1 increments. The degree of cooperation was measured as the amount of money the participant decide to give. This was a "one-shot" interaction, meaning that it only took place once.

Procedure

Participants were informed that the study required English as the first language, that it was voluntary to partake, and that they had the right to withdraw at any time. Further, they were informed that the experiment involved interacting with another person regarding money, and this description was aimed to be as neutral as possible by avoiding more value-laden words such as being cooperative or about winning or losing. Participants were paid \$0.5 to complete the study and told that they would keep any money from the interaction task.

After having agreed to partake in the study, participants were informed about how the interaction task (i.e., the GSDG) works and that both parties had to submit their response before they were shown the other's response. To ensure that the participants had understood the task, they had to answer hypothetical questions correctly before being able to continue (see **Supplementary Material Appendix I**). Subsequently, participants were presented with the demographic questions, followed by the well-being questions. Participants were randomly assigned to either answer the word-response questions in random order first or the rating scales in random order first.

Before the interaction task started, participants were presented with a message reading, "Searching for another person. Please wait," and after 16 s another sign popped up reading, "Connecting you with another person." Participants were then presented with a summary of the instructions of the game and the response alternatives regarding the amount to give to the other person. When they had answered, the participants were presented with the text reading, "Please wait while processing. The other person cannot see your response." This was followed by the message: "Please wait for the other person to submit their decision." Subsequently, they were presented with the result of the task (e.g., "The other person decided to give you \$0. You gave \$0. In total, you get \$1 and the other person gets \$1.").

After the interaction, participants answered two questions about their momentary experience of harmony in life and

satisfaction with life (which were not analyzed or reported in this study due to its exploratory nature) followed by the TDM. Lastly, before being debriefed, the participants were asked to leave feedback about the interaction. The study took on average 16 min to complete.

Ethical Considerations

The studies received ethical approval from the Regional Ethical Committee in Lund, Sweden. Prior to participating, all participants received information about the study and were asked for consent to participate. They were informed that participation was anonymous and voluntary and that they could withdraw at any time without having to give a reason. At the end of the study, the participants were given more information about the study and were informed about the deception and why it was important, and they were informed that because of this deception they received the maximum possible amount from the GSDG.

Natural Language Processing and Statistical Analyses

The Semantic Space and Representations

The word data were analyzed with the r-package Text 0.9.0[2] (Kjell et al., 2021). The words generated in the current study were given their semantic representations (i.e., vectors of numeric values describing each word) from a previously created semantic space (used and described in Kjell et al., 2021). The semantic space was created using latent semantic analyses (Landauer and Dumais, 1997) based on singular values decomposition (Golub and Kahan, 1965) on the co-occurrences of 1.7×10^9 words from the English Google 5-g database. The semantic space includes semantic representations for the 120,000 most frequent English words, in which each word is described in 512 dimensions (for more details, see Kjell et al., 2016).

Word responses were cleaned in accordance to the procedures put forward in Kjell et al. (2019). Words were spelled according to American spelling, and misspelled words were corrected only when the meaning was clear, otherwise they were ignored. Successively repeated words or instances of "NA" or similar were removed. Answers comprising sentences or strings of words rather than one descriptive word in each response box were removed. And words that did not have a semantic representation in the semantic space were returned as missing values.

Because the responses to the word-response questions comprised several words, the semantic representations of the words were added together using the mean of each dimension to create one representative semantic representation for each word-response question. These semantic representations were subsequently used to create *semantic similarity scales,* language predicted *valence*, and the *word plots* as specified below.

Semantic Similarity Scales (SSS)

The values that compose the semantic representations can be seen as coordinates in a high-dimensional space, and the closer together the semantic representations of two words/texts are the more semantically similar they are. Hence, the semantic similarity between two words/texts can be represented by the cosine of the angle between the two semantic representations (Landauer and Dumais, 1997). In the current study, we measured the level of a psychological construct by measuring the semantic similarity between responses to the word-response questions and the corresponding word-norms. For example, if a person's response to the harmony in life question was semantically similar to the harmony in life word-norm, this person was considered to have a high level of overall harmony in life. High unipolar semantic similarity is the semantic similarity to the targeted construct (e.g., harmony in life), low unipolar semantic similarity scales are the opposite of the target constructs (e.g., disharmony in life), and bipolar semantic similarity scales are the low unipolar scale subtracted from the high unipolar scale (e.g., the harmony in life SSS minus the disharmony in life SSS).

Language Predicted Scales

The values in the semantic representations can also be used in multiple regressions to create models predicting certain semantic characteristics of a word/text. In the current study, we employed language predicted valence scales. These are based on the ANEW word list where approximately 1,000 words have been rated by individuals in terms of their negative or positive valence. In the multiple regression ($y = \underline{c}^* \underline{x}$), the semantic representations ($\underline{x}$; i.e., vectors) of the words were used to predict the valence (y) rated by participants, in which the coefficient ($\underline{c}$) describes the relationship between the words and the valence. This regression model was applied to the word responses in the current study to estimate their valence (i.e., the regression model was a language predicted valence scale). This model was created using ridge regression (with a penalty grid ranging from 10^{-16} to 10^{16}), where cross-validation was used to evaluate the model (for more details, see Kjell et al., 2021).

The SSS and the language predicted scales were used in the correlations to understand their relationship to rating scales and cooperation.

Supervised Dimension Projection Plots

Plots were used to visualize words that were statistically significant in relation to the specified categories or dimensions (i.e., axes) under investigation. In the current study words that significantly differed in their semantic representation between responses to the harmony in life versus the satisfaction with life questions were plotted on the x-axis, and on the y-axis the words were plotted according to the degree of cooperation. Words that statistically significantly differed on a specified dimension, were plotted in color (rather than in gray), and the font size of the word indicated its frequency in the data.

The supervised dimension projection plot compares two groups' responses to different questions (e.g., harmony in life versus satisfaction with life responses) or low versus high cooperation on a scale using mean split. To achieve this, a semantic representation is first constructed to capture the difference between the two groups, and this semantic representation (point in space) can be seen to form a line through the origo (and is referred to as *the aggregated direction embedding line*). The *aggregated direction embedding* is constructed by taking

[2]www.r-text.org

the mean of all semantic representations in each group and then subtracting the two representations.

Finally, all the individual words in the word responses are "projected" onto the *aggregated direction embedding* line. The projection is achieved by first "anchoring" all of the individual words' representations in space by subtracting the second group's aggregated semantic representation from each word's representation and then using the dot product to project each word's anchored representation (for more details, see Kjell et al., 2021). To statistically test the words, a dot product null distribution is created by calculating the dot product among randomly selected semantic representations and an *aggregated direction embedding* created from randomly swapping words' semantic representations from the two groups. Multiple comparisons are corrected using the false discovery rate (FDR) correction.

Statistical Analyses

To examine the relationships between variables, Pearson *r* are used when both variables are normally distributed, and Spearman's *rho* are used when at least one of the variables are not normally distributed. To examine the relationship between two variables whilst controlling for other variables we use partial correlation (e.g., see Kim, 2015).

R-Packages

All analyses were carried out in R (R Core Team, 2020) using RStudio (RStudio Team, 2020). Apart from the text package (Kjell et al., 2021), the following packages were used: tidyverse (Henry and Wickham, 2020), Hmisc (Harrell et al., 2020), dplyr (Wickham et al., 2020), ppcor (Kim, 2015), psychometric (Fletcher, 2010), reshape2 (Wickham, 2007), ggplot2 (Wickham, 2016, p. 2), data.table (Dowle and Srinivasan, 2019), lm.beta (Behrendt, 2014), lattice (Sarkar, 2008), effsize (Torchiano, 2020), and WRS2 (Mair and Wilcox, 2019).

RESULTS

Descriptive Statistics

Ninety-six participants (53%) were categorized as prosocials, 70 (39%) were categorized as proselfs, and 15 (8%) were uncategorized. On average the participants gave \$0.45 (SD = 0.41; prosocials: Mean = \$0.52, SD = 0.41; proselfs: Mean = \$0.34, SD = 0.39). The cooperation variable exhibited a bimodal, rather than a normal, distribution, and the semantic similarity scales contained some considerable outliers. Because a few participants had, for example, just replied *yes* or *no* to the word-response questions, and because both of these opposing answers yielded outliers of low semantic similarity, outliers with a z-score more extreme than ± 3.29 were removed for all semantic similarity scales (see **Table 1**). **Table 2** presents correlations among the included well-being measures. The highest correlation was between the HILS and SWLS ($r = 0.84$, $p < 0.001$), whereas the computational language assessments showed lower intercorrelations (e.g., the semantic similarity score of the harmony in life responses and norms with the satisfaction with life responses and norms yielded an r of 0.59, $p < 0.001$).

The Well-Being and Cooperation Objective

In accordance with H_1, the CLA of overall harmony in life (i.e., the SSS between the word-responses of the harmony question and the harmony in life word-norm) was positively correlated with cooperation, and this was strongest in prosocials ($r = 0.35$, $p < 0.001$; see **Table 3**). However, in contrast to H_1, this relationship was not found with the HILS. In contrast to H_2, measures of overall satisfaction with life were not significantly related to cooperation. **Figure 1** shows these correlations, where the correlations were controlled for age, gender, perceived financial situation, and all the other well-being-related measures (all presented in **Table 3**), and only the correlation between the computational language assessment of overall harmony in life and cooperation was significant ($r = 0.41$, $p < 0.001$). It is also worth noting that there is a significant positive correlation between the Disharmony semantic similarity scale and cooperation among proselfs ($r = 0.39$, $p < 0.001$).

The Methodological Objective

In support of H_3, the distinct prediction of cooperation was shown with computational language assessments but not with numerical rating scales. The computational language assessment of harmony in life also clearly supported the prediction in H_1, but this was not the case for the HILS. However, in relation to H_2 there were no significant correlations among the satisfaction with life measures and cooperation (see **Figure 1**).

The Computational Language Assessment-Based Plot

Figure 2 shows the statistically significant word responses according to the type of open-ended question (x-axis) and to the level of cooperation (y-axis). In regard to H_4, the relationships hypothesized in H_1 and H_2 were observed considering that there were more words that were significantly more closely related to high harmony in life that were also significantly related to a high level of cooperation, as compared with high satisfaction with life. That is, 10 words are significant in the right upper corner (see legend; including *peace*, *happiness*, *balance*, *harmony*, and *unity*) whereas there are 0 significant words in the right lower corner. On the other side, there are only 2 words (*fulfilled* and *annoyed*) related to overall satisfaction with life and high cooperation, but 4 words related to satisfaction with life and low cooperation (including *happy*, *proud*, *unhappy*, and *satisfied*).

DISCUSSION

Well-Being and Cooperation Objective

We have demonstrated a clear link between subjective well-being and cooperation. Specifically, the computational language assessment of harmony in life yielded a moderately strong significant positive correlation with degree of cooperation in

TABLE 1 | The number of participants excluding missing values, the range, the mean, and the standard deviation before and after outliers were removed for each variable.

Measure	*N*	Range	Mean	*SD*
HILS	181	5–35	26.5	6.34
SWLS	181	5–35	23.9	7.58
H-LPV	180	3.28–7.84	6.07	0.97
S-LPV	178	3.25–7.63	5.98	1.00
H-SSS	180	−0.03–0.72	0.33	0.16
S-SSS	178	0.04–0.72	0.33	0.14
Dh-SSS	179	0.01–0.36	0.16	0.08
Ds-SSS	178	0.03–0.64	0.27	0.10
Ds-SSS no outliers[1]	177	0.03–0.56	0.26	0.09

N, number of participants excluding missing values; SD, standard deviation. HILS, Harmony in Life Scale; SWLS, Satisfaction with Life Scale; H, harmony; S, satisfaction; Dh, disharmony; Ds, dissatisfaction; LPV, language predicted valence; SSS, Semantic Similarity Scale.
[1]Only the Ds-SSS variable included outliers.

TABLE 2 | Pearson correlations among the wellbeing-related measures.

	1	2	3	4	5	6	7	8	9	10
(1) HILS										
(2) SWLS	0.84***									
(3) H-LPV	0.67***	0.61***								
(4) S-LPV	0.65***	0.64***	0.72***							
(5) H-SSS	0.45***	0.42***	0.71***	0.60***						
(6) S-SSS	0.48***	0.51***	0.52***	0.76***	0.59***					
(7) Dh-SSS	−0.24**	−0.18*	−0.18*	−0.03	0.06	0.12				
(8) Ds-SSS	−0.54***	−0.50***	−0.51***	−0.54***	−0.35***	−0.11	0.23**			
(9) H-Dh-SSS	0.54***	0.48***	0.73***	0.56***	0.88***	0.47***	−0.43***	−0.42***		
(10) S-Ds-SSS	0.66***	0.67***	0.69***	0.89***	0.65***	0.85***	−0.02	−0.62***	0.60***	

*N = 177–181; *p < 0.05, **p < 0.01, ***p < 0.001. HILS, Harmony in Life Scale; SWLS, Satisfaction with Life Scale; H, harmony; S, satisfaction; Dh, disharmony; Ds, dissatisfaction; LPV, language predicted valence; SSS, Semantic Similarity Scale.*

TABLE 3 | Spearman's *rho* for self-reports and cooperation for the various groups.

Social value orientation	HILS	SWLS	H- LPV	S- LPV	H-SSS	S-SSS	Dh-SSS	Ds-SSS	H-Dh-SSS	S-Ds-SSS
All (*N* = 181)	0.06	0.09	0.12	0.05	0.18*	0.10	0.27***	0.02	0.05	0.10
Prosocials (*n* = 96)	0.04	0.02	0.21*	0.17	0.35***	0.16	0.17	−0.08	0.25*	0.21*
Proselfs (*n* = 70)	−0.06	−0.07	−0.04	−0.15	−0.09	−0.08	0.39***	0.08	−0.28*	−0.14

**p < 0.05, ***p < 0.001. HILS, Harmony in Life Scale; SWLS, Satisfaction with Life Scale; H, harmony; S, satisfaction; Dh, disharmony; Ds, dissatisfaction; LPV, language predicted valence; SSS, Semantic Similarity Scale.*

prosocials, while the computational language assessment of overall satisfaction with life did not. This held true even when controlling for all other studied well-being measures (including the traditional numeric rating scales and the predicted valence of the word responses), gender, age, and perceived financial situation.

The word figures further support the importance of harmony in life in relation to cooperation. The statistically different word responses between the harmony in life and satisfaction with life are consistent with previous research; for example, *peaceful* and *calm* are related to harmony in life, and *happy* and *fulfilled* are related to satisfaction with life (Kjell et al., 2019). Importantly, the *Cooperation*-axis further supports that overall harmony in life, but not overall satisfaction with life, is positively related to cooperation, considering that words such as *peace*, *balance*, *harmony*, and *unity* are related to both high harmony in life and cooperation, whereas words such as *happy*, *proud*, and *satisfied* are related to satisfaction with life responses and low levels of cooperation.

Different conditions and situations that support and promote cooperation have been extensively researched (see e.g., Calcott, 2008). Cooperation is a particularly integral part of human society, where human cooperation can be attributed to well-developed *cognitive* resources (Stevens and Hauser, 2004).

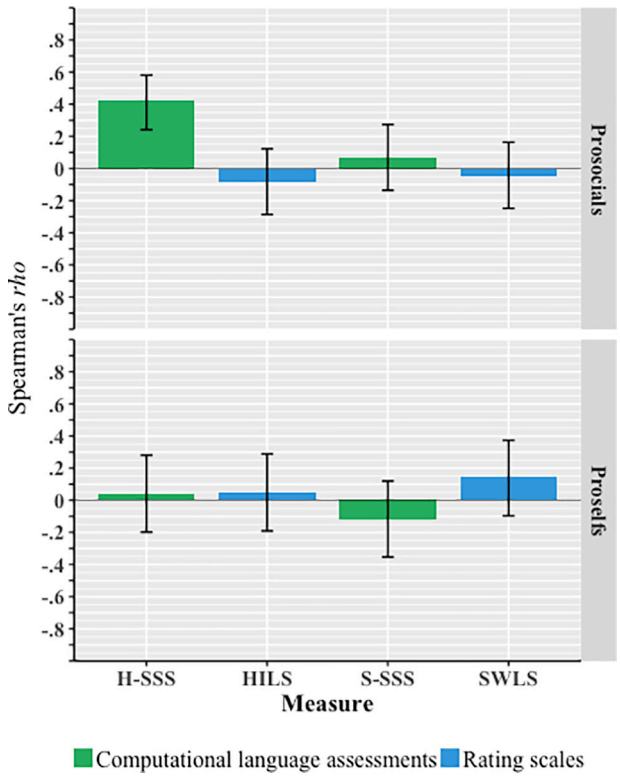


FIGURE 1 | Partial Spearman's *rho* correlation (with 95% confidence interval) between each well-being measure and cooperation. The data were controlled for the three remaining well-being measures, language predicted valence, perceived financial situation, gender, and age for the various groups. Only H-SSS for prosocials was significant ($r = 0.41$; $p < 0.001$); $n = 69$ prosocials, $n = 59$ proselfs. SSS, Semantic Similarity Scale; H, harmony; S, satisfaction; HILS, Harmony in Life Scale; SWLS, Satisfaction with Life Scale. Note that the *n* differs from **Table 3** because partial correlation requires no missing values on all variables.

However, previous research has particularly examined whether certain *emotions* (e.g., Ketelaar and Au, 2003; de Hooge et al., 2007; Nelissen et al., 2007; Kjell and Thompson, 2013) or *positive mood* (Proto et al., 2019) lead to increased cooperation, and less focus has been put on the cognitive component of subjective well-being. To our knowledge, this is the first experiment that tests and demonstrates an association between cooperation and harmony in life measured as the cognitive component of subjective well-being.

Considering the importance of cooperation for societies, we believe that the current results warrant further research interest to deepen the understanding of the link to harmony in life and to satisfaction with life. The results may be seen as particularly important for the subjective well-being literature because there currently is a rather narrow understanding of well-being that predominantly focuses on satisfaction with life. This relates to Kjell's (2011) concern that a one-sided satisfaction with life focus:

> "Appears likely to encourage the individual to put themselves and their expectations first rather than allowing for an adaptive balance of both satisfaction and balance/harmony. Furthermore, measuring satisfaction while neglecting balance/harmony, might crucially relate to the issue that one person's satisfaction can result in another person's dissatisfaction." (p. 260).

Thus, overall, the results give support to the concerns that an overemphasized focus on satisfaction with life can be considered to one-sidedly reflect self-regard and self-centeredness (e.g., see Christopher, 1999; Kjell, 2011), and they suggest that harmony in life is important in complementing satisfaction with life within the subjective well-being approach (see also Kjell et al., 2016).

Prosocials and Proselfs

Whereas there was a positive correlation between *harmony* semantic similarity scores and cooperation among prosocials as expected; the results revealed a positive correlation between *disharmony* semantic similarity scores and cooperation among proselfs. That is, among proselfs higher levels of cooperation was related to higher semantic similarity between their harmony in life word-responses and the *disharmony* word norm (i.e., a negative valenced word norm). This finding may perhaps be compared with how inducing proselfs with *guilt* (i.e., a negative valenced emotion) increases their cooperation (Ketelaar and Au, 2003; de Hooge et al., 2007). However, to further understand this relationship among proselfs require further research.

The Methodological Objective

From a methodological perspective, this study shows that open-ended, computational language assessments of well-being are distinctly related to a theoretically relevant behavioral outcome, whereas standard, closed-ended numerical rating scales are not. As previously discussed, these differences are also discerned in the word figures, where the rating scales method lack an equivalent descriptive analytic method (since rating scales do not allow for descriptive word responses).

To our knowledge this is the first research study that supports the validity of self-reported computational language assessments over traditional rating scales in relation to actual behaviors. Research has, for example, shown that self-reported computational language assessments demonstrate very high convergence with rating scales (Kjell et al., 2021) and that computational language assessments yield higher validity in categorizing external stimuli, including facial expressions (Kjell et al., 2019). There is also evidence that computational language assessments based on individuals' social media texts (rather than question-based, prompted, self-reports) can predict personality (Schwartz et al., 2013) and are correlated with depression in medical records (Eichstaedt et al., 2018).

Thus, the results presented here add to the research literature demonstrating the validity of computational language assessments. We suggest that future research should attempt to identify the boundary conditions of the computational language assessments (e.g., identifying conditions when ratings scales may have higher validity than computational language assessments and where a combination might be preferred). It would also be valuable to examine respondents' preferences for the different response formats. For example, which format do respondents

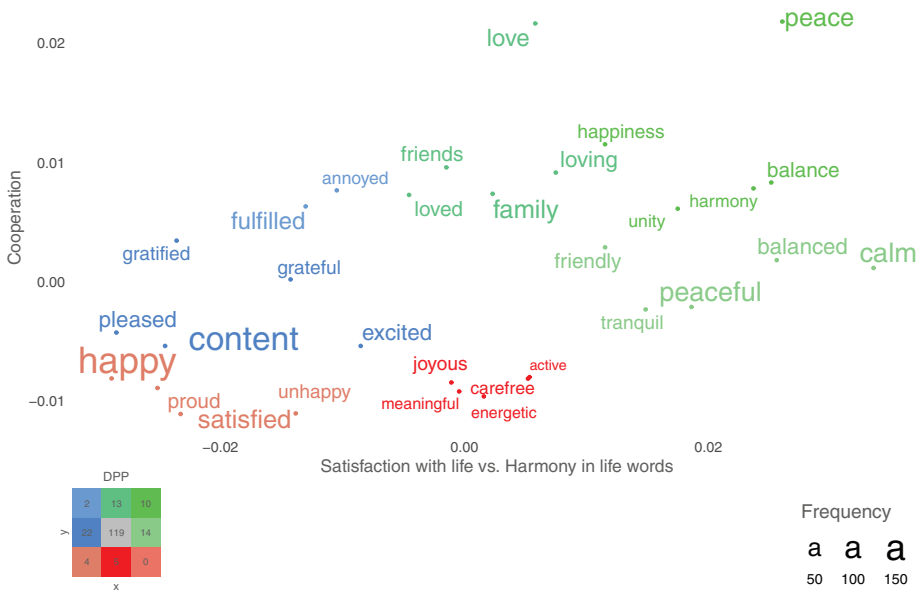


FIGURE 2 | A supervised dimension projection plot of words significantly differing between responses to the satisfaction with life and the harmony in life (*x*-axis) and the level of cooperation (*y*-axis). The colored legend in the lower left corner indicates the color and number of significant words in each part of the figure (for example, there are 10 light green words that are significantly high on both the *x*-axis and *y*-axis). *N* = 180.

prefer in regard to how easy it is to use or how well they can describe their mental states.

Limitations

The current study has some limitations. It examined only a specific type of cooperation that was constrained to one interaction with an "anonymous" person about money, and participants only received the extreme amounts (i.e., all or nothing). Future studies could also examine harmony versus satisfaction in social dilemmas that, for example, include repeated interactions concerning more aspects than just money. In addition to replicating the current results, future research could examine whether the cooperative link between well-being and cooperation differs in different contexts and situations.

Buhrmester et al. (2011) demonstrated that using Mechanical Turk to collect data produces comparable results as more conventional and standard methods, while also ensuring good generalisability. However, future studies could examine these effects when participants are recruited from other, more social contexts. Further, the analyses statistically controlled for several factors, including perceived financial situation and other well-being measures; however, to further our understanding of the computational language assessments, future studies could control for participants' current emotional state as well as personality traits. Lastly, the current study did not record the time required to answer the different assessment methods. Whereas Kjell et al. (2019) found that it took longer time for participants to answer the open-ended word format than the rating scales format when describing facial expressions; they also found that only using one rather than ten descriptive words when describing their own mental health produced reliable, although somewhat less accurate, predictions. Future studies could examine how many responses that are necessary while preserving high validity and reliability, how long time each method take to complete and whether respondents prefer one assessment method over the other.

CONCLUSION

From a methodological perspective, the results support the validity of computational language assessments, and computational language assessments can distinctly reveal the theoretically relevant behavioral outcome of cooperation within a social dilemma game in relation to subjective well-being, while traditional rating scales cannot. From a well-being perspective, the results reveal a distinct behavioral difference between harmony in life and satisfaction with life, with harmony in life being to a higher degree related to cooperative behavior.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**. Data will be made available at https://osf.io/bqnar/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Regionala etikprövningsnämnden i Lund. The

patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed to the study design. OK and SS performed the data collection and the natural language processing analyses. OK, DD, and SS were involved in the other analyses as well as writing up the manuscript. All authors approved the final version of the manuscript for submission.

FUNDING

This research was supported by The Swedish Research Council (ID: 2019-06305) funded an international postdoc for OK. Lund University Library funded the cost for open access.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.601679/full#supplementary-material

REFERENCES

Balliet, D., Parks, C., and Joireman, J. (2009). Social Value Orientation and Cooperation in Social Dilemmas: a Meta-Analysis. *Group Process. Intergroup Relat.* 12, 533–547. doi: 10.1177/1368430209105040

Behrendt, S. (2014). *Lm.Beta: Add Standardized Regression Coefficients To Lm-Objects*. URL: https://CRAN.R-project.org/package=lm.beta.

Bradley, M. M., and Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology*, Florida: University of Florida.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspect. Psychol. Sci.* 6, 3–5. doi: 10.1177/1745691610393980

Calcott, B. (2008). The other cooperation problem: generating benefit. *Biol. Philos.* 23, 179–203. doi: 10.1007/s10539-007-9095-5

Christopher, J. C. (1999). Situating Psychological Well-Being: exploring the Cultural Roots of Its Theory and Research. *J. Couns. Dev.* 77, 141–152.

de Hooge, I. E., Zeelenberg, M., and Breugelmans, S. M. (2007). Moral sentiments and cooperation: differential influences of shame and guilt. *Cogn. Emot.* 21, 1025–1042. doi: 10.1080/02699930600980874

Diener, E., Emmons, R. A., Larsen, R. J., and Griffin, S. (1985). The satisfaction with life scale. *J. Pers. Assess.* 49, 71–75.

Dowle, M., and Srinivasan, A. (2019). *Data.Table: Extension Of 'Data.Frame'*. URL: https://CRAN.R-project.org/package=data.table.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., et al. (2018). Facebook language predicts depression in medical records. *Proc. Natl. Acad. Sci. U. S. A.* 115, 11203–11208.

Fletcher, T. D. (2010). *Psychometric: Applied Psychometric Theory*. URL: https://CRAN.R-project.org/package=psychometric.

Golub, G., and Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *J. SIAM Numer. Anal.* 2, 205–224.

Harrell, F. E. Jr., and with contributions from Charles Dupont and many others. (2020). *Hmisc: Harrell Miscellaneous*. URL: https://CRAN.R-project.org/package=Hmisc.

Henry, L., and Wickham, H. (2020). *Rlang: Functions For Base Types And Core R And "Tidyverse" Features*. URL: https://CRAN.R-project.org/package=rlang.

Ketelaar, T., and Au, W. T. (2003). The effects of feelings of guilt on the behaviour of uncooperative individuals in repeated social bargaining games: an affect-as-information interpretation of the role of emotion in social interaction. *Cogn. Emot.* 17, 429–453. doi: 10.1080/02699930143000662

Kim, S. (2015). *Ppcor: Partial And Semi-Partial (Part) Correlation*. URL: https://CRAN.R-project.org/package=ppcor.

Kjell, O. N., and Diener, E. (2020). Abbreviated three-item versions of the satisfaction with life scale and the harmony in life scale yield as strong psychometric properties as the original scales. *J. Pers. Assess.* 103, 183–194.

Kjell, O. N. E. (2011). Sustainable Well-Being: a Potential Synergy Between Sustainability and Well-Being Research. *Rev. Gen. Psychol.* 15, 255–266. doi: 10.1037/a0024603

Kjell, O. N. E., Daukantaitė, D., Hefferon, K., and Sikström, S. (2016). The Harmony in Life Scale Complements the Satisfaction with Life Scale: expanding the Conceptualization of the Cognitive Component of Subjective Well-Being. *Soc. Indic. Res.* 126, 893–919. doi: 10.1007/s11205-015-0903-z

Kjell, O. N. E., Giorgi, S., and Schwartz, H. A. (2021). Text: an R-package for analyzing and visualizing human language using natural language processing and deep learning. *PsyArXiv Preprints*. doi: 10.31234/osf.io/293kt

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Kjell, O. N. E., and Thompson, S. (2013). Exploring the impact of positive and negative emotions on cooperative behaviour in a Prisoner's Dilemma Game. *PeerJ* 1:e231. doi: 10.7717/peerj.231

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295x.104.2.211

Li, C. (2006). The confucian ideal of harmony. *Philos. East West* 56, 583–603. doi: 10.1353/pew.2006.0055

Li, C. (2008). The Philosophy of Harmony in Classical Confucianism. *Philos. Compass* 3, 423–435.

Mair, P., and Wilcox, R. (2019). Robust statistical methods in R using the WRS2 package. *Behav. Res. Methods* 1–25.

Mason, W., and Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behav. Res. Methods* 44, 1–23.

McClintock, C. G. (1972). Social motivation: a set of propositions. *Behav. Sci.* 17, 438–455. doi: 10.1002/bs.3830170505

Nelissen, R. M. A., Dijker, A. J. M., and deVries, N. K. (2007). How to turn a hawk into a dove and vice versa: interactions between emotions and goals in a give-some dilemma game. *J. Exp. Soc. Psychol.* 43, 280–286. doi: 10.1016/j.jesp.2006.01.009

Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* 45, 867–872. doi: 10.1016/j.jesp.2009.03.009

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgm. Decis. Mak.* 5, 411–419.

Proto, E., Sgroi, D., and Nazneen, M. (2019). Happiness, cooperation and language. *J. Econ. Behav. Organ.* 168, 209–228.

R Core Team (2020). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. Switzerland: R Core Team.

RStudio Team (2020). *RStudio: Integrated Development Environment for R. RStudio, PBC*. Switzerland: RStudio Team .

Ryan, R. M., and Deci, E. L. (2001). On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annu. Rev. Psychol.* 52, 141–166.

Sarkar, D. (2008). *Lattice: Multivariate Data Visualization with R*. Germany: Springer.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., et al. (2013). Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One* 8:e73791. doi: 10.1371/journal.pone.0073791

Stevens, J. R., and Hauser, M. D. (2004). Why be nice? Psychological constraints on the evolution of cooperation. *Trends Cogn. Sci.* 8, 60–65. doi: 10.1016/j.tics.2003.12.003

Torchiano, M. (2020). *effsize: Efficient Effect Size Computatio. R package version 0.8.1n*.

Van Lange, P. A. M., and Kuhlman, D. M. (1994). Social value orientations and impressions of partner's honesty and intelligence: a test of the might versus morality effect. *J. Pers. Soc. Psychol.* 67, 126–141. doi: 10.1037/0022-3514.67.1.126

VanLange, P. A. M., Otten, W., DeBruin, E. M. N., and Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *J. Pers. Soc. Psychol.* 73, 733–746.

Wickham, H. (2007). Reshaping Data with the reshape Package. *J. Stat. Softw.* 21, 1–20.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wickham, H., François, R., Henry, L., and Müller, K. (2020). *Dplyr: A Grammar Of Data Manipulation*. URL: https://CRAN.R-project.org/package=dplyr.

Conflict of Interest: OK and SS have co-founded WordDiagnostics, which uses Computational Language Assessments for diagnosing mental health issues.

The remaining 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.

frontiers in Psychology

ORIGINAL RESEARCH
published: 12 May 2021
doi: 10.3389/fpsyg.2021.604135

Reevaluating the Influence of Leaders Under Proportional Representation: Quantitative Analysis of Text in an Electoral Experiment

Annika Fredén[1] *and Sverker Sikström*[2]*

[1] Department of Political, Historical, Religious and Cultural Studies, Karlstad University, Karlstad, Sweden, [2] Department of Psychology, Lund University, Lund, Sweden

OPEN ACCESS

Edited by:
Kim F. Nimon,
University of Texas at Tyler,
United States

Reviewed by:
Kathrin J. Hanek,
University of Dayton, United States
Jan Ketil Arnulf,
BI Norwegian Business School,
Norway
Oyvind Lund Martinsen,
BI Norwegian Business School,
Norway

***Correspondence:**
Sverker Sikström
sverker.sikstrom@psy.lu.se

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

Received: 09 September 2020
Accepted: 14 April 2021
Published: 12 May 2021

Citation:
Fredén A and Sikström S (2021) Reevaluating the Influence of Leaders Under Proportional Representation: Quantitative Analysis of Text in an Electoral Experiment. Front. Psychol. 12:604135. doi: 10.3389/fpsyg.2021.604135

We propose that leaders play a more important role in voters' party sympathy in proportional representation systems (PR) than previous research has suggested. Voters, from the 2018 Swedish General Election, were in an experiment asked to describe leaders and parties with three indicative keywords. Statistical models were conducted on these text data to predict their vote choice. The results show that despite that the voters vote for a party, the descriptions of leaders predicted vote choice to a similar extent as descriptions of parties. However, the order of the questions mattered, so that the first questions were more predictive than the second question. These analyses indicate that voters tend to conflate characteristics of leaders with their parties during election campaigns, and that leaders are a more important aspect of voting under PR than previous literature has suggested. Overall, this suggests that statistical analysis of words sheds new light of underlying sympathies related to voting.

Keywords: leaders, parties, voting, primacy, proportional representation, statistical semantics

INTRODUCTION

Most current election studies measure political sympathy through approval rating scales (see for example, Mueller, 1970; Van der Eijk and Franklin, 2009; Oscarsson and Holmberg, 2013). However, a measure like a score on a scale tells little about the contents of the voter's evaluation. What role does the leader play? How much relates to policy? This study approaches party preference from a new angle, asking voters directly what they think about when they think about parties, and to what extent leaders intertwine with descriptions of the party. The focus of the study is proportional representation systems (PR), where electoral studies tend to center on ideology, parties and party identification rather than leaders (Granberg and Holmberg, 1988; McCall Rosenbluth and Shapiro, 2018; Oscarsson and Holmberg, 2020). Nevertheless, the party leaders should be important as spokespersons and concrete representations of policy orientation, especially in a political landscape where many voters switch parties from one election to the next (Fieldhouse et al., 2020). This study argues for the inclusion of leader perceptions in studying voters' behavior, also under proportional representation.

The focus on voters' own responses in the present study is rather unique: so far, the materials that are the focus in related studies are usually party manifestos, press releases or related materials (Klüver and Sagarzazu, 2016; Crabtree et al., 2018). When leaders are the focus, the current trend is survey experiments where leader qualities are experimentally manipulated (see for example Tavares et al., 2018). Fewer studies refer to "real" political leaders, which is the starting-point in this study. Media scholars have been somewhat more tempted to follow this path where, for example, Aaldering et al. (2018) start from the perspective that the tone of the media coverage of leaders has a mediating impact on the propensity to vote for a party. Still, current research tends to look at leader impact during election campaigns more generally, without asking the voters themselves.

We collected voters' free text descriptions in a real-life election campaign—the 2018 Swedish General Election. The party system contains a large number of smaller parties, which makes it possible to examine the influence of leaders for those too. In order to emphasize the party vs. the leader in the experiment, half of the sample was randomly assigned to describe the leaders first, whereas the other half started by describing the parties. Drawing on findings from the psychology literature (Murdock, 1962; Sullivan, 2019), the belief was that a primacy effect should make a statement that comes first matter more for the voting decision than a statement that comes after, independently of whether it concerns the party, or the leader.

These claims are supported by the following hypotheses. In current media, the party leader is the concrete representation of the abstract concept of a party. Because concrete and simple representations are usually easier to understand and remember (see e.g., Kahneman, 2011), the hypothesis is that the leader representation will be essential for shaping the voter's associations to a party. At the same time, party policies are important shortcuts for orienting oneself in a party system with a clear left-right ideological spectrum. The argument is that voters under proportional representation can have difficulties separating leaders' policy messages from their parties, and parties from their leaders. Citizens thus need both representations: the concrete of the leader, and the more stable ideological reference to the party, to form an association of a political unit. This leads to the first hypothesis:

Leader conflation hypothesis (H1). *The words a voter uses to describe a party leader tend to be similar and are at least as indicative for his or her vote choice as the words used to describe the party.*

The second hypothesis concerns how the order of the descriptive task potentially affects the predictive powers of free text descriptions. A well-studied effect in the memory literature is the primacy effect (e.g., Murdock, 1962). This effect shows that items that are presented first are usually better remembered than items presented later. The theoretical basis for the primacy effect is not fully understood, however, a view typically taken in the literature relates to the first items receives more attention or are rehearsed more than the later items (Anderson and Hubert, 1963; Sullivan, 2019). More important for the present study, is that text written early tends to carries more important content. In particular, Kjell et al. (2019) showed a *semantic primacy effect*, where words generated early in the description of a mental state were more predictive of rating scale scores, than words generated later. This finding matches the current experiment well, in the sense that the descriptions that voters give first should be more strongly associated with vote intention than the descriptions that they give later. The words that the voter comes up with first are the words that are most easily accessible, and represent the voter's primary view of a political unit (i.e., the mental representation of the party and/or the leader), whereas words that generated later are less informative the voter's representation of the political unit. Following this line of argument we propose that:

Primacy hypothesis (H2). *In the condition where voters are asked to first describe leaders and then describe parties, the description of leaders will be a more important indicator of vote choice than the description of parties. The opposite pattern will be found in the conditions were voters are asked to describe parties first.*

From these perspectives, the overarching expectation is that voters' descriptions of leaders during election times are equally important for their choices as their descriptions of the parties. Their respective predictive powers will also depend on the order of the descriptive task, since more important, concrete and consistent descriptions should be remembered earlier.

The text descriptions were analyzed using latent semantics, which is a natural language processing (NLP) approach to quantitative text (Landauer and Dumais, 1997) which we combine with machine learning (ML) to predict voting behavior. This method allows examination of how respondents' descriptions of parties and leaders co-occurred, and how these descriptions can be related to vote choice. In line with the argument, the descriptive words of leaders and parties predicted vote choice to the same extent, whereas the order of questions mattered. The words that the respondent gave first predicted the vote intention better that the words that came second.

MATERIALS AND METHODS

The case for the study is the proportional representation system of Sweden, which was long dominated by the single party Social Democrats governments. More recently it has oriented toward coalitions of parties (Bäck and Bergman, 2016; Fredén, 2021). The party system of 2018 consisted of three bigger parties (the Social Democrats, the Moderates, and the Sweden Democrats) and five smaller parties (Greens, Liberals, Left party, Centre party, and Christian Democrats). The focus of the present study is the parties that characterize these types of PR systems, namely, these smaller parties. One circumstance that could direct voters more toward leaders over parties in general is if the parties coordinate before the election, or if the parties run more independently. If the negotiations between the parties after the election are supposed to matter more, that is, if the blocs are more loosely organized, then, candidate evaluations potentially matter more

since the leaders will then have a crucial role in the post-election negotiations. In the 2018 general election, the parties competed more independently than in the previous elections (Aylott and Bolin, 2019). Three of the parties had new party leaders since the previous election (the Greens, the Christian Democrats, and the Moderates), and three of the parties were at risk of not reaching the four percent electoral threshold (the Greens, the Christian Democrats, and the Liberals). The presence of a strengthened populist party, the Sweden Democrats, oriented the campaign toward issues as well as the four percent electoral threshold, since the established blocs needed the smaller parties to reach the threshold to survive as government alternatives. One of the main opinion polls indicated a tight race between the three bigger parties Social Democrats, Moderates, and Sweden Democrats (Bergman, 2018) and most polls suggested a close race between the traditional left-socialist bloc and the center-of-right bloc (see for example Sifo, 2018).

Study Design

The aim of the study was to collect evaluations of political parties and their leaders in a real-life campaign using a survey experimental design, where we would (1) examine voters' leader descriptions in relation to their party descriptions (2) examine the impact of priming the respondent with the leader descriptive task vs. the party descriptive task. The experiment was part of a methods-oriented survey at the Swedish National Election Studies Program/LORe Internet Campaign panel, managed by the SOM-Institute, University of Gothenburg. It was released 2 weeks before the general election on 9 September 2018 (respondents continued to submit their responses up to the Election Day, but most of the respondents submitted their answers in the period 25–31 August). Before entering the study, participants agreed to participate by accepting the data and investigation procedures in the LORe Internet campaign panel, in accordance with current ethics and GDPR standards.

Sample

The sample consists of self-recruited participants, who participated in the survey voluntary (with no extra reward). Eleven thousand six hundred twenty-one were invited to take the survey experiment, and 58% (6,776) responded. Mullinix et al. (2015) show that convenience samples, in general, generate effects that are very similar to population-based samples. Since the main interest here is the global relationship between party and leader perceptions, rather than contents, levels of support or word counts concerning specific parties, sample characteristics should matter relatively little (compare Mutz et al., 2019). The number of unique words is high: 10,010 related to parties and 8,165 related to leaders. Most important, standard socio-economic characteristics are evenly spread between the randomized treatment groups. Respondents come from all age groups, education levels and gender (for more detailed information of sample characteristics, see the **Supplementary Table A1**). Also party support is evenly spread between the two treatment groups. Supporters of the main parties Social Democrats and Moderates are underrepresented compared with election results, whereas supporters of smaller parties are overrepresented (compare Valmyndigheten, 2019, and **Supplementary Table A2** for distributions of vote intentions over treatments in this experiment). Seventy-nine percent of the respondents indicated that they were very certain about their party choice when they took the survey experiment (corresponding to 6 or 7 on a scale from 1 to 7, where 1 stands for not certain at all, and 7 for absolutely certain). The study sample is thus a group of relatively convinced voters. Since the impact of leaders on choice may be stronger among volatile and unknowledgeable voters (Oscarsson and Holmberg, 2016), the potential leader influence on the perception of a political party should not be particularly great here. Instead, the experiment should rather underestimate than overestimate primacy effects and leader conflation.

Selection of Political Parties

For pragmatic reasons, we had to select a smaller number of parties to include for the descriptive task in the experiment. Including too many parties in the survey experiment would also have made the task more cumbersome and risk increasing participant fatigue. Previous political science research mainly focuses on leader effects of bigger parties (compare research from the US context as well as previous research on the Swedish context such as for example Oscarsson and Holmberg, 2016). Here, the focus is on party characteristics that are typical for proportional representation, that is, smaller parties whose fortune is more insecure during elections times, and where the leader may play a less salient role. The survey includes the three smallest parties that were at risk of not reaching electoral representation—the Liberals, the Greens, and the Christian Democrats—and the major right-wing party, the Moderates, which was a potential leader of government. This implies a mix of parties in terms of size, their positions on the left-right-scale, as well as the gender of the leader (two male leaders, and two female leaders). In order to draw conclusions about potential leader effects under proportional representation, this sample of four parties should thus serve as a relevant reference[1].

Experimental Procedure

The online survey experiment proceeded as follows. Participants were randomly assigned to starting with either the task of describing parties (n = 3,428), or the task describing the leaders of the same four parties (n = 3,348). The party item was formulated as follows "What does the following party represent for you?" "Please enter up to three descriptive keywords, or leave blank if you do not know about the party." The party leader item, in turn, was formulated as follows: "What does the following party leaders/spokespersons represent for you?" "Please enter up to three descriptive keywords, or leave blank if you do not know about the party leader/spokesperson." Respondents were provided with party abbreviations in brackets when they were

[1]The Social Democrats, the largest party in Sweden, was not included in the survey. However, their main competitor, the Moderates, is included. Since the previous dominance of Social Democrats in Sweden is an exception rather than rule in similar PR contexts (see, for example, McCall Rosenbluth and Shapiro, 2018), this sample should still be sufficiently representative for established proportional representation party systems today.

to describe the leaders. Since this is how leaders are usually presented in the media, we believe that is a valid way of collecting words on leaders. See the **Supplementary Material** for the original formulations in Swedish.

On the next page, the descriptive task was shifted—those who had not described leaders described parties, and vice versa[2]. All respondents described all four parties and their four party leaders. The survey institute decided the order in which the parties appeared[3]. On the following screen, the respondent indicated three important issues. After these items, the respondent declared his or her vote intention. The experiment finished by responding to a question about certainty of vote decision on a scale from1 (not certain at all) to 7 (very certain).

For screenshots of the experiment's online format, see the **Supplementary Material**.

Method: Latent Semantic Analysis Predicting Voting From Text Data

The novelty of this study is to collect free text descriptions of political units (leaders and parties), as a complement to the standard approval rating scales. The primary interest was to study how well these three keywords generated by the participants predict their voting behavior, and to what extent priming respondents with one descriptive task over the other would influence the results. To our knowledge, the best methods for doing this builds on a combination of NLP and ML. NLP methods allow quantification of texts (e.g., keywords) to a high dimensional representation to which an individual's word descriptions are compared. ML allows us to investigate whether this representation predicts an outcome variable, which in our case is voting behavior. To do this, we used Latent Semantic Analysis (LSA), a quantitative text analytical approach that quantifies and systematize voters' responses. This data-driven (unsupervised) method is suitable for measuring meaning in word expressions by quantifying how similar the words are to each other. The method resembles factor analysis, since words that are similar in meaning receives similar semantic representations. In this study, we first created a high dimensional (N = 300) semantic representation based on the 135,806 words in the dataset. A semantic space was created based on the words generated by the participants. The method is described in detail in Kjell et al. (2019) (see also Landauer and Dumais, 1997). First a word-by-word co-occurrence matrix is created where each cell represents the number of times two words have been generated in the same answer by a participant. Then each cell is normalized by logarithm plus one. Finally, a data compression algorithm (singular value decomposition) is applied to this matrix, where the first 300 resulting dimensions are maintained (i.e., the dimensions are ordered by how much information they maintain from the original matrix, so the first dimensions are the most important)[4]. This results in a representation where each word is associated with a vector (normalized to the length of one) that represent how semantically similar the words are in the dataset. Since the data material concerns keywords on parties and politicians (and little irrelevant text information) this method is suitable for categorizing responses. The three words from the individual are summarized to one semantic representation, by adding the vector associated to each word and normalize the length of the resulting vector to one. This representation allows to measure the semantic similarities scores between two texts, as well as make predictions to a numerical variable, for example vote intention, as described below.

The semantic *similarity score* (SS) between two sets of words is calculated by taking the cosine of the angle between two associated semantic vectors, which in this case is mathematically equivalent with multiplying each dimension with each other and summing them. This score, bounded between -1 and +1, which is high when the word sets are similar in meanings and small when they are unrelated. For example, descriptions such as "right" get a high score relative to a "conservative" dimension, since these are close in meanings, whereas descriptions such as for example "solidarity" gets a lower score relative to an "authority" dimension, since these word representations are less similar to each other.

The semantic representations of the three words that the participants generated can be used to *predict* vote choice. This is done by using the semantic representation as predictors in logistic regression, where 1 represent choosing the specific party, 0 choosing some other party. The resulting predicted vote choice were then correlated with the empirical value of the vote intention. Here we used point-biserial correlation, which is a suitable method for dichotomous dependent variables[5]. Based on the text data from a specific word question (e.g., about the party "Moderates"), we can predict to what extent voters are likely to choose a party (e.g., "Moderates," "Liberals," etc.). For example, if Liberal party voters tended to enter the words "liberal" and "school" together and other voters used these word combinations to a less extent (or used words with very different meanings), such systematic co-occurrence patterns will translate into r-scores that are higher for the Liberal party relative to other parties. The predictions are evaluated with a 10-fold cross-validation procedure, which means that the text data from the experiment was randomly divided into a training set consisting of 90% of the data, were the empirical values of vote intentions were

[2]The average number of words that the respondent used for describing parties and leaders decreased for parties and leaders depending on whether it was the first or last descriptive task. It decreases somewhat more for parties (from on average 2.3 to 1.9 words) than for leaders (from 2.0 to 1.8 words).

[3]The descriptive statistics show that the average number of words is very similar describing the four leaders, independent of their internal order in the survey experiment (ranging from 1.96 to 2.03 when the leader descriptive task comes first, to 1.80–1.87 when it comes last). When it comes to the party descriptive task, the number of keywords is associated with size rather than order: the greatest number of words (2.46) is for the Moderate party and the lowest (2.17) for the Christian Democrats when party descriptions came first, and 2.05 for Moderates (highest) vs. 1.80 for Christian Democrats (lowest) when party descriptions came last. Since the general patterns are similar and the differences relatively small, we do not believe that an internal order effect is driving the results and main conclusions. Moreover, previous research indicates that the first and second words are most important for measuring the respondent's semantic representation (Kjell et al., 2019).

[4]Another option would have been to compare the voters' text descriptions with general text materials such as for example google n-grams, however, in this case, there was reason to create a semantic space of words related to political parties and leader descriptions.

[5]See for example Medium (2019) for a discussion on correlation measures for binary outcomes.

used in the predictions, and then evaluated on the remaining 10% of the data. This procedure is repeated 10 times, with different training and test data sets, so all data points receive a predicted value. The **Supplementary Material** provides a general overview of this method.

We thus predicted vote choice based on the survey items that contained up to six words per political unit (three related to the party, three related to the party leader) and the vote intention item, which were collected during the experiment. This allowed direct comparison of predictive powers of words related to leaders, vs. words related to parties (H1). We separated the sample into test order, where one condition consisted of respondents answering the party leader questions first, and the other condition answered the party items first. This allowed us to investigate whether test order influenced the results (H2).[6]

To get a qualitative overview of the data, the words in the dataset were also visualized in word clouds, following the methods specified in Kjell et al. (2019). The words that were representative for voters' descriptions of leaders and parties were grouped together, where the words in the center of the clouds are the most representative (i.e., words with the highest semantic similarity with other words in the same condition), and font size represents frequency. Then, these descriptions were divided by order, i.e., coming first or last as descriptive tasks.

The analyses were performed in the Matlab version of the online statistical software semanticexcel.com (Sikström et al., 2020).

RESULTS

Descriptives

First, we evaluate leader and party descriptions depending on the order of the question. **Figure 1** summarizes descriptions of all four leaders, where the left side of the figure shows words that are indicative results of the leader question being second (i.e., after the party question), whereas the right-hand side presents the result when the leader question was presented first. When the party leaders were described first, the descriptions relate to politics and party characteristics, for example "school" [skola], "conservative" [konservativ], as well as personal characteristics, such as for example "boring" [tråkig]. On the other hand, when leaders were described after the parties, the word clouds contain less ideological and issue-related words, and more characteristics related to personal qualities: "trustworthy" [trovärdig], "competent" [duktig]. These findings give some first support to the hypothesis that the leader and party descriptions tend to conflate, especially if the leader item precedes the party item.

For comparison, **Figure 2**, in turn, shows word clouds for the party descriptions, where the left-hand side shows words indicative of party descriptions given first, and the right side party descriptions after leader descriptions. Interestingly, we find that the most central word is identical to the most central leader descriptions that come first: "school" [skola]. In addition, more abstract concepts such as "freedom" and the "EU" are significant in the party descriptions that precede leader descriptions. The interpretation of the difference between words coming first or last is less straight-forward for parties than for leaders. The size of the cloud, i.e., the number of central words following the LSA, is the same size in the two treatments. One observation is that the words on the right, i.e., where the party descriptions come last, are more influenced by policy-laden words (for example, "right" [höger]), which are features that may detach voters from a party. It is possible that the leader descriptions that preceded these descriptions influenced the party words in that direction.

The descriptions suggest that participants describe leaders and parties with rather similar concepts if it is their first associative task. Nevertheless, personal characteristics such as "boring" [tråkig] and "clear"/"unclear" [tydlig/otydlig] are significant words following the first descriptive leader task. This suggests that primacy of leaders can influence voters to think about issues and personal characteristics simultaneously, and that evaluations of leaders and party contents in conjunction predict vote choice to the greatest extent.

Correlations

Below we test the hypotheses more directly, i.e., how well the written descriptions of leaders and parties predicted voting intention. **Table 1** and **Figure 3** show the point biserial correlation (r) between the empirical value of vote intention and the predicted value of vote choice. **Table 1** shows how well descriptions of the party's leader or party predicted vote choice for the four parties that were included in the survey items (the Moderates, the Liberals, the Christian Democrats, and the Greens). These analyses support the first hypothesis that voters' descriptions of leaders are associated with vote choice to the same extent as their description of parties. Overall, leader descriptions (r = 0.125, s = 0.0093) mattered as much as party descriptions (r = 0.127, s = 0.0093) concerning these four focal parties.

Second, we find support of the primacy effect stated in the second hypothesis. The descriptions that the voters gave first, in general, predicted vote choice better independently of the descriptive task. Thus, for example, if leaders were described first, then the descriptions of these predicted vote choice better than the descriptions of the parties that came afterward. The first question had a higher correlation for parties (r = 0.145, s = 0.013 vs. r = 0.110, s = 0.013) as well as for party leaders (r = 0.149, s = 0.013 vs. r = 0.101, s = 0.013). The correlation for the first questions were significantly higher than the correlation for second questions (p = 0.0026, N = 2,607 (participant) * 8 (questions), z = 3.0 (see Meng et al., 1992)[7].

Figure 3 illustrates the general pattern that we found. The graph compares the predictive powers of vote choice at t1 (when the party or leader is described first) and at t2 (when party or

[6]The sample sizes in the sixteen different correlation models (the word descriptions of four parties and their leaders described first or last, correlated with respondents' vote intentions) varied between 2,475 and 2,844.

[7]The only exception is the Moderate party descriptions, where the second party descriptive task predicted vote choice to a greater extent that the first party descriptive task. One potential explanation is that the preceding leader descriptions amplified voters' associations to the party more in this case because the leader is a potential Prime Minister.

FIGURE 1 | Word descriptions of leaders under different conditions. The figures show words arranged in word clouds. The number of plotted words has been limited to 100. The total number of words for the leader descriptions is 53,372, and the number of unique words is 8,165. Words in color were significant following Bonferroni correction for multiple comparisons. The figure shows color-coded data-points that significantly discriminate between the high and the low value of the scale parties first **(left)** or leaders first **(right)** using semantic tests. The semantic t-test comparing the two sets of leader descriptive texts was significant $t(53{,}372) = 23.69$, $p = 0.0000$. For a detailed description of the method, see the **Supplementary Figure notes**.

FIGURE 2 | Word descriptions of parties under different conditions. The figures show words arranged in word clouds. The number of plotted words has been limited to 100. The total number of words for the party descriptions is 59,515, and the number of unique words is 10,010. Words in color were significant following Bonferroni correction for multiple comparisons. The figure shows color-coded data-points that significantly discriminate between the high and the low value of the scale parties first **(left)** or leaders first **(right)** using semantic tests. The semantic t-test comparing the two sets of party descriptive texts was significant $t(59{,}515) = 23.61$, $p = 0.0000$. For a detailed description of the method, see the **Supplementary Figure notes**.

leader is described last). This supports the conclusion that the order of the descriptive task matters (Hypothesis 2).

To summarize, the latent semantic analyses support the claim that voters' descriptions of leaders and parties are of similar importance for predicting their vote choice. In line with our first hypothesis, the leader descriptions from the three keywords predicted vote intention to the same extent as party descriptions did. Leader descriptions given before party descriptions were more influential and explicitly related to policy. This suggests that voters often conflate representations of leaders and parties, and that these concepts may be exchanged in the voter's mental representation within the context of voting behavior. In addition, the generally clearer descriptions that voters entered in the first party association task appear to matter more for choice than the more diverging words that summarized the last descriptive task. Thus, the more solid picture of the party and its leader predicted vote choice better than the less coherent figure. Nevertheless, the analysis shows that the leader descriptions, which are more oriented toward evaluation of personal qualities, can be part of this solid conceptualization of the party. In our experiment, we find that associations that are prior to others predict vote choice best, which demonstrates that a primacy effect occurs in the vote decision-making process.

DISCUSSION

The results from an electoral experiment and a LSA lent support to the hypothesis that descriptions of leaders had about equally as strong predictive power as descriptions of parties in the 2018 Swedish general election campaign. We also found clear evidence that the order of the questions matter: descriptions of leaders or parties that were given first mattered more for the decision and were qualitatively different from descriptions given second. We thus revealed a primacy effect in an electoral context, where voters were asked to describe party leaders and parties in free text. One potential implication is that the piece of information that the campaign currently emphasizes, be it the leader or the party, is influencing the voter's mindset. The analysis also

TABLE 1 | Prediction of voting intention based on participants written descriptions.

	Describe parties			Describe leaders		
Party	**First**	**Second**	**Leader**	**First**	**Second**	**Average**
Green	0.104	0.070	I. Lövin	0.144	0.131	0.112
Liberals	0.186	0.059	J. Björklund	0.153	0.084	0.121
Christian Democrats	0.073	0.060	E. Busch Thor	0.087	0.024	0.061
Moderates	0.215	0.252	U. Kristersson	0.211	0.164	0.211
Average	0.145	0.110		0.149	0.101	0.126

The sample sizes in the sixteen different correlation are based on the samples of the respective treatments group (n = 3,428 when parties are described first, n = 3,348 when leaders are described first). The models include respondents with valid key word responses. The actual sample sizes of the prediction models vary between 2,844 (Moderate party described first) and 2,475 (Green's spokesperson I. Lövin described second).

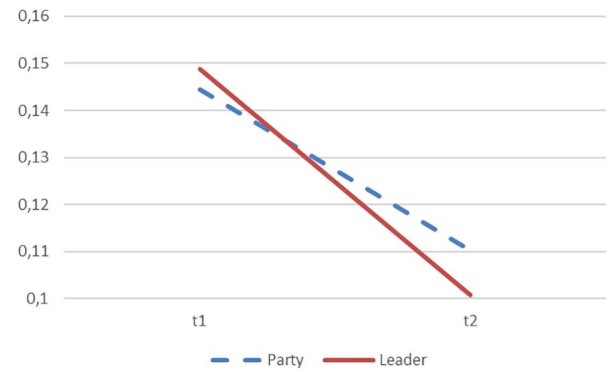


FIGURE 3 | Prediction of voting intention based on order of the descriptive task. The figure is based on **Table 1** and show the party average Pearson correlation coefficient between predicted and empirical voting intention (r) at t1 (party or leader described first) and t2 (party or leader described last). The difference between the correlations at t1 and t2 are significant with $p < 0.001$.

showed that a combination of policy and personal characteristics had greater predictive power than personal characteristics that are less associated with the party. Studying voters' own free text responses thus revealed that leader influence on political sympathy is salient also in PR.

Using this kind of text analytical approach advances knowledge about how voters think when they think about parties and leaders, and how these associations guide the vote choice process. This knowledge may have practical implications, as it suggests that creating positive associations to the leader and make them stand in the front of the party's policy message is a potentially successful party strategy. Leader and party descriptions are not separate from policy positions, and the leader's role as spokespersons should not be underestimated. Clarity and uniqueness in the policy message, as well as repetition of it, would make such associative patterns even more salient. The influence of leaders can be a problem if this has consequences for party survival that are not rooted in policy responsiveness between voters and parties, but rather in personal characteristics of the leader that can be less relevant. Nevertheless, if the parties' paint a coherent picture of party policies and leader, it will facilitate voters' possibility to predict the leaders' forthcoming abilities to negotiate with other parties. In the studied election, previous policy orientations had to be reconsidered since the election resulted in unclear majorities. Future studies should look deeper into which part influences the other most during the election campaign: i.e., if parties and leaders can influence voters directly through emphasizing certain dimensions in their repertoires (compare Broockman and Butler, 2017; Barber and Pope, 2018) or whether these associations rather grow from "below," i.e., the voters.

Forthcoming studies should also elaborate more upon how important leaders are for party success, and how important leaders are as spokespersons for certain policy profiles. For example, the choice of leader has an impact on how voters view the party's ideological leaning, which in turn affects voting behavior. When voters tend to be more volatile, and rely upon various media sources for their decisions, these kinds of mechanisms become even more important to scrutinize. One avenue for future research is the duration of such leader effects, and the potential variation over different contexts. This study examined a proportional context with a less predictable outcome than usual as a populist party had grown stronger relative to the established parties. Potentially, this made the 2018 Swedish election more similar to other countries where we have seen similar patterns, such as Denmark, Norway, and the United Kingdom. It would be fruitful to replicate the study in these other contexts in order to test the generalizability of the relatively strong leader influence we found in this experiment.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the datafiles may still contain personal identifiable information. Parts of the dataset may be available on upon request, after some additional screening by data managers at the Laboratory of Opinion Research at the University of Gothenburg. Requests to access the datasets should be directed to AF, annika.freden@kau.se.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University of Gothenburg. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AF and SS developed the study concept and raised the funds that were necessary to conduct the experiment (with AF as main applicant). AF was responsible for the final survey experimental design, the contact with the Lore opinion lab at the University of Gothenburg, performed the statistical overview analysis of the

experimental data, developed a strategy for the more complex analysis in collaboration with SS, and drafted the manuscript. SS performed the statistical analysis and provided important revisions. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Vetenskapsrådet Grant 2017-02941. Publication fees were supported by Lund University.

ACKNOWLEDGEMENTS

The authors would like to thank the reviewers and research assistant Alexander Rangfält.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.604135/full#supplementary-material

REFERENCES

Aaldering, L., Van der Meer, T., and Van der Brug, W. (2018). Mediated leader effects: the impact of newspapers' portrayal of party leadership on electoral support. *Int. J. Press Polit.* 23, 70–94. doi: 10.1177/1940161217740696

Anderson, N. H., and Hubert, S. (1963). Effects of concomitant verbal recall on order effects in personality impression formation. *J. Verbal Learn. Verbal Behav.* 2, 379–391. doi: 10.1016/S0022-5371(63)80039-0

Aylott, N., and Bolin, N. (2019). A party system in flux: the Swedish parliamentary election of September 2018. *West Eur. Polit.* 42, 1504–1515. doi: 10.1080/01402382.2019.1583885

Bäck, H., and Bergman, T. (2016). "The parties in government formation," in *The Oxford Handbook of Swedish Politics*, ed. J. Pierre (Oxford: Oxford University Press), 206–226.

Barber, M., and Pope, J. S. (2018). does party trump ideology? disentangling party and ideology in America. *Am. Polit. Sci. Rev.* 113, 38–54. doi: 10.1017/S0003055418000795

Bergman, L. (2018). *Yougov: Sista Mätningen Inför Valet - Så Ligger Partierna Till*. Available online at: https://www.metro.se/artikel/yougov-sista-m%C3%A4tningen-inf%C3%B6r-valet-s%C3%A5-ligger-partierna-till [Accessed February 7 2019]

Broockman, D. E., and Butler, D. M. (2017). The causal effects of elite position-taking on voter attitudes: field experiments with elite communication. *Am. J. Polit. Sci.* 61, 208–221.

Crabtree, C., Golder, M., Gschwend, T., and Indridason, I. H. (2018). *It's Not Only What you Say, It's Also How You Say It: The Strategic Use of Campaign Sentiment*. Ann Arbor, MI: University of Michigan, doi: 10.31235/osf.io/g2sd6

Fieldhouse, E., Green, J., Evans, G., Mellon, J., Prosser, C., Schmitt, H., et al. (2020). *Electoral Shocks. The Volatile Voter in a Turbulent World*. Oxford: Oxford University Press, doi: 10.1093/oso/9780198800583.001.0001

Fredén, A. (2021). How polling trends influence compensational coalition-voting. *Front. Polit. Sci.* doi: 10.3389/fpos.2021.598771 [Epub ahead of print].

Granberg, D., and Holmberg, S. (1988). *The Political System Matters: Social Psychology and Voting Behaviour in Sweden and the United States*. Cambridge: Cambridge University Press.

Kahneman, D. (2011). *Thinking, Fast and Slow*. London: Penguin Books Ltd.

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115. doi: 10.1037/met0000191

Klüver, H., and Sagarzazu, I. (2016). Setting the agenda or responding to voters? political parties, voters and issue attention. *West Eur. Politics* 39, 380–398. doi: 10.1080/01402382.2015.1101295

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295X.104.2.211

McCall Rosenbluth, F., and Shapiro, I. (2018). *Responsible Parties: Saving Democracy From Itself*. New Haven, CT: Yale University Press.

Medium. (2019). *An Overview of Correlation Measures Between Categorical and Continuous Variables*. Available at https://medium.com/@outside2SDs/an-overview-of-correlation-measures-between-categorical-and-continuous-variables-4c7f85610365 [Accessed September 17 2019].

Meng, X., Rubin, D. B., and Rosenthal, R. (1992). Comparing correlated correlation coefficients. *Psychol. Bull.* 111, 172–175. doi: 10.1037/0033-2909.111.1.172

Mueller, J. E. (1970). Presidential popularity from Truman to Johnson. *Am. Polit. Sci. Rev.* 64, 18–34. doi: 10.2307/1955610

Mullinix, K. J., Leeper, T. J., Druckman, J., and Freese, J. (2015). The generalizability of survey experiments. *J. Exp. Polit. Sci.* 2, 109–138. doi: 10.1017/XPS.2015.19

Murdock, B. (1962). Serial position effect of free recall. *J. Exp. Psychol.* 64, 482–488. doi: 10.1037/h0045106

Mutz, D. C., Pemantle, R., and Pham, P. (2019). The perils of balance testing in experimental design: messy analyses of clean data. *Am. Statist.* 73, 32–42. doi: 10.1080/00031305.2017.1322143

Oscarsson, H., and Holmberg, S. (2013). "Party leader effects on the vote," in *Political Leaders and Democratic Elections*, eds K. Aarts, A. Blais, and H. Schmitt (Oxford: Oxford University Press), 35–51.

Oscarsson, H., and Holmberg, S. (2016). *Swedish Voters. [Svenska Väljare.]*. Stockholm: Wolters Kluwer.

Oscarsson, H., and Holmberg, S. (2020). *The Research Handbook on Political Partisanship*. Cheltenham, UK: Edward Elgar Publishing.

Sifo (2018). Available online at: https://www.kantarsifo.se/rapporter-undersokningar/valjarbarometern [Accessed September 23 2019]

Sikström, S., Kjell, O. N. E., and Kjell, K. (2020). SemanticExcel.com: an online software for statistical analyses of text data based on natural language processing. *Statist. Semant. Methods Appl. Springer Int. Publis.* 2020, 87–103. doi: 10.1007/978-3-030-37250-7_6

Sullivan, J. (2019). The primacy effect in impression formation: some replications and extensions. *Soc. Psychol. Personal. Sci.* 10, 432–439. doi: 10.1177/1948550618771003

Tavares, G. M., Sobral, F., Goldszmidt, R., and Araújo, F. (2018). Opening the implicit leadership theories' black box: an experimental approach with conjoint analysis. *Front. Psychol.* 9:100. doi: 10.3389/fpsyg.2018.00100

Valmyndigheten (2019). https://www.val.se/valresultat/riksdag-landsting-och-kommun/2018/valresultat.html [Accessed September 23 2019]

Van der Eijk, C., and Franklin, M. N. (2009). *Elections and Voters*. Basingstoke: Palgrave Macmillan.

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

frontiers in Psychology

ORIGINAL RESEARCH
published: 04 June 2021
doi: 10.3389/fpsyg.2021.602581

Freely Generated Word Responses Analyzed With Artificial Intelligence Predict Self-Reported Symptoms of Depression, Anxiety, and Worry

Katarina Kjell, Per Johnsson and Sverker Sikström*

Department of Psychology, Lund University, Lund, Sweden

OPEN ACCESS

Edited by:
Kim F. Nimon,
University of Texas at Tyler,
United States

Reviewed by:
Merylin Monaro,
University of Padua, Italy
Jan Ketil Arnulf,
BI Norwegian Business School,
Norway

****Correspondence:***
Katarina Kjell
katarina.kjell@psy.lu.se

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

Received: *03 September 2020*
Accepted: *11 March 2021*
Published: *04 June 2021*

Citation:
Kjell K, Johnsson P and Sikström S (2021) Freely Generated Word Responses Analyzed With Artificial Intelligence Predict Self-Reported Symptoms of Depression, Anxiety, and Worry. Front. Psychol. 12:602581. doi: 10.3389/fpsyg.2021.602581

Background: Question-based computational language assessments (QCLA) of mental health, based on self-reported and freely generated word responses and analyzed with artificial intelligence, is a potential complement to rating scales for identifying mental health issues. This study aimed to examine to what extent this method captures items related to the primary and secondary symptoms associated with Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD) described in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). We investigated whether the word responses that participants generated contained information of all, or some, of the criteria that define MDD and GAD using symptom-based rating scales that are commonly used in clinical research and practices.

Method: Participants (N = 411) described their mental health with freely generated words and rating scales relating to depression and worry/anxiety. Word responses were quantified and analyzed using natural language processing and machine learning.

Results: The QCLA correlated significantly with the individual items connected to the DSM-5 diagnostic criteria of MDD (PHQ-9; Pearson's r = 0.30–0.60, $p < 0.001$) and GAD (GAD-7; Pearson's r = 0.41–0.52, $p < 0.001$; PSWQ-8; Spearman's r = 0.52–0.63, $p < 0.001$) for respective rating scales. Items measuring primary criteria (cognitive and emotional aspects) yielded higher predictability than secondary criteria (behavioral aspects).

Conclusion: Together these results suggest that QCLA may be able to complement rating scales in measuring mental health in clinical settings. The approach carries the potential to personalize assessments and contributes to the ongoing discussion regarding the diagnostic heterogeneity of depression.

Keywords: diagnostic criteria, major depressive disorder, generalized anxiety disorder, measurement, artificial intelligence, natural language processing, machine learning, diagnostic assessment

INTRODUCTION

Closed-ended rating scales are commonly used in clinical practice and research to assess the type and severity of mental health issues [e.g., the Patient Health Questionnaire-9 (PHQ-9); Kroenke et al., 2001, and the Generalised Anxiety Disorder Scale-7 (GAD-7); Spitzer et al., 2006]. These rating scales require the respondent to rate their agreement with predefined items designed to target the construct/disorder being measured.

Question-based computational language assessment (QCLA) is an alternative method to rating scales (Kjell et al., 2019). This method has an open-ended word-response format that allows the respondent to freely elaborate on their state of mind using descriptive words or texts that are analyzed using natural language processing and machine learning. Previous research shows that QCLA *measures*, *describes*, and *differentiates* well between psychological constructs (Kjell et al., 2019) when compared with the total score of rating scales specifically designed to capture the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria (American Psychological Association [APA], 2013). This study aimed to further investigate the QCLA method by examining to what extent it captures individual items related to the primary and secondary symptoms associated with mental health aspects described in the DSM using the PHQ-9 (Kroenke et al., 2001) and the GAD-7 (Spitzer et al., 2006). These rating scales are designed to target the DSM criteria for Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD; American Psychological Association [APA], 2013).

Computational Language Assessment

Computational language assessments have been used to predict and monitor depression on a population level using *naturally occurring text* on social media (e.g., Mowery et al., 2016). Posts on social media have also been used to predict further episodes of depression several months before onset using Twitter (De Choudhury et al., 2013; Reece et al., 2017) and Facebook (e.g., Eichstaedt et al., 2018). Eichstaedt et al. (2018) predicted depression as recorded in participants' medical records using language from Facebook posts. They further predicted episodes of depression 3 months before they were documented in the medical records, suggesting that prediction models based on social media might be a useful complement in diagnostic screening procedures (Eichstaedt et al., 2018). However, less research has been done when it comes to QCLA where participants are asked about aspects of their mental health.

Question-Based Computational Language Assessment

Kjell et al. (2019) constructed QCLAs with the aim of measuring and describing mental health, including depression, worry, harmony in life, and satisfaction with life, and evidence suggests that this method quantifies constructs with similar or greater validity compared with traditional rating scales. For example, in two studies participants were asked to describe facial expressions from a validated database, and it was found that QCLAs accurately categorized significantly more facial expressions compared with rating scales. It was further demonstrated that QCLAs of subjective experience predict rating scales' total score with correlations of $r = 0.58$ for the GAD-7, $r = 0.59$ for the PHQ-9, $r = 0.72$ for the Harmony in life scale, and $r = 0.63$ for the Satisfaction with life scale ($p < 0.001$ for all r-values; $N = 477$). In another study, it was demonstrated that the QCLA of harmony in life was significantly correlated with cooperative behavior ($r = 0.18$ for all participants and $r = 0.35$ for participants categorized as prosocial); whereas the corresponding rating scale of harmony in life (Kjell et al., 2021a) did not demonstrate a significant correlation (Kjell et al., 2021a).

The QCLA Method

The word or text responses generated from questions on QCLAs are suitable for statistical analyses based on the creation of high-dimensional word embeddings from a large language corpus. The QCLA method quantifies words, or texts, as a vector, i.e., the word embedding of word responses, and uses this word embedding to construct three types of measures: *semantic similarity scales*, *language-trained scales,* and *language-predicted valence scales* (Kjell et al., 2019).

Semantic similarity scale

The semantic similarity scale has the advantage of being able to measure a construct based on an empirically generated semantic definition of a construct. This is achieved by creating a word norm (a list of empirically derived words) that participants have generated to describe the construct being measured. The semantic similarity scale is measured by the semantic similarity (closeness) between the participants' word responses and the targeted word norm. This procedure is carried out without any involvement of a rating scale.

Word norms may describe the two endpoints of a psychological construct e.g., "being depressed" or "not at all depressed" (Kjell et al., 2020). A unipolar semantic similarity scale is the semantic similarity between word responses and *one* word norm (e.g., depression responses and the being depressed word norm). A bipolar semantic similarity scale is the semantic similarity between the text generated from the to-be-measured questions and the similarity scores between the subtraction of two word norms (e.g., depression responses to the being depressed word norm minus the depression responses to the not at all being depressed word norm). Unipolar and bipolar semantic similarity scores for depression and worry have been found to correlate well with the total scores of rating scales, where bipolar scores correlate stronger than unipolar scores (Kjell et al., 2020).

Language-trained scale

The language-trained scale can be used to measure a construct using word or text data by linking it to well-established rating scales or other quantifiable markers related to the concept. For example, the word embedding generated by a word-response question about depression can be used to predict the rating scale score of the PHQ-9. This prediction can be done, for example, by using multiple linear regression or other machine

learning methods, and the validity can be evaluated with cross-validation methods. The measure of accuracy of the cross-validated predictions is calculated by the correlation between the predicted and actual scores.

Language-predicted valence scale

It is possible to take a prediction model trained on one dataset and apply it to another dataset. Valence is an important dimension on which emotions can be described and identified (e.g., Bradley and Lang, 1999). Kjell et al. (2019) trained a valence model using the Affective Norms for English Words, where participants have rated the emotional valence ranging from unpleasant to pleasant of more than 1,000 words (ANEW; Bradley and Lang, 1999). The model was then used to predict the valence scores of word responses, where the predicted valence scores were found to be strongly correlated with rating scale scores.

Semantic Similarity Scales Versus Language-Trained Scales

A language-trained scale can be trained to estimate a rating scale. However, this scale is typically constructed with different items, where some items can be predicted with higher accuracy than other items. In contrast, semantic similarity scales rely on the agreement between how respondents answering the word-response questions and how respondents creating the word norm understand the construct being measured. Thus, there is a fundamental difference between these that we investigate further.

QCLA in the Clinical Setting

A response format where a person describes their mental health with their own words has many potential advantages in clinical practice. For example, it allows for patient-centered care because it focuses on the patient's unique set of symptoms. Further, QCLA may add knowledge for a more patient-centered approach to routine outcome measures used in everyday clinical practice to monitor symptom severity and treatment effectiveness (e.g., see de Beurs et al., 2011; Washington and Lipstein, 2011). QCLA may also identify co-occurring symptoms that otherwise would have been undetected or may increase awareness of domains that are important to patients but that are not targeted, or captured, in rating scales.

Measuring Psychiatric Disorders Versus Subjective Well-Being

The open-ended nature of the QCLA method taps into an interesting difference between the measurement of psychiatric disorders versus subjective well-being. Assessments of psychiatric disorders (e.g., DSM-5, ICD-10) are strictly criteria driven, whereas measures of subjective well-being aim to be criteria free (e.g., Kesebir and Diener, 2008; Kjell and Diener, 2020). Corresponding to this, the open-ended word response format of word-response questions enables respondents to express themselves more freely than when responding to closed-ended rating scales. In contrast, diagnosing individuals with psychiatric disorders involves assessing whether individuals fulfill a specific set of diagnostic criteria stated in manuals such as the DSM-5 (American Psychological Association [APA], 2013). Hence, these approaches differ in whether it is the patients/clients or the professionals who specify the evaluation criteria. In subjective well-being measures, it is the respondent who is assumed to be best suited to judge their level of well-being (Kesebir and Diener, 2008); whereas for psychiatric disorders it is a trained researchers or mental health care professionals who define if someone meets the specified diagnostic criteria (American Psychological Association [APA], 2013). This difference makes it important to understand the word-responses' relationship to individual rating scales' items of criteria-based mental health disorders.

DSM Criteria

The DSM-5 categorizes the criteria into *primary* and *secondary* for both MDD and GAD. For an individual to be diagnosed with MDD, they have to meet one of the primary criteria and five in total (including the primary and secondary criteria). To be diagnosed with GAD an individual has to meet all of the primary criteria and at least three of the secondary criteria. The rating scales used in this study are explicitly designed to capture the symptoms and criteria outlined in the DSM (e.g., for depression this includes disabilities in areas such as sleep, concentration, and movement; Kroenke et al., 2001). On the other hand, the QCLA questions only focus on assessing the respondent's own understanding of depression (e.g., *Over the last 2 weeks, have you been depressed or not?*) and does not explicitly probe about specific symptoms that respondents not explicitly report following the question about being depressed. Thus, indirect symptoms such as changes in sleep, concentration, or movements can, but do not necessarily need to be reported. Hence, it is important to understand to what extent the broad question currently used in QCLA also captures specific symptoms and to what extent it might be necessary to also ask respondents specific symptom-related word-response questions.

Major Depressive Disorder

The two *primary* DSM-5 criteria for MDD focus on the subjective experiences of depression (i.e., depressed mood and loss of interest), and the *secondary* criteria focus on related symptoms such as psychomotor agitation or retardation, diminished ability to concentrate, and weight loss/gain. Hence, the primary symptoms are arguably closer to how individuals primarily think about depression, whereas individuals might not directly associate the secondary criteria as strongly with being depressed.

Reviewing the PHQ-9 shows that the nine items capture each of the DSM-5 diagnostic criteria well (see **Table 1**). In contrast, the QCLA for depression captures individuals' subjective experiences and their own understanding of depression, which potentially might be more related to the primary rather than the secondary criteria. That is, instructions for the QCLA for depression (and worry) are broad and generally stated: "*Write descriptive words relating to those aspects that are most important and meaningful to you*" (see the section "Materials and Methods" for full details; Kjell et al., 2019).

Generalised Anxiety Disorder

The two *primary* DSM-5 criteria for GAD focus on the experience of excessive worry and having difficulties in controlling one's worrying, whereas the *secondary* criteria mainly focus on related symptoms such as muscle tension, irritability, and fatigue. Also,

TABLE 1 | DSM-5 diagnostic criteria for Major Depressive Disorder.

DSM-5 criteria	Primary	PHQ-9 items
Five or more symptoms including depressed mood and/or loss of interest or pleasure, during a 2-week period.		
Depressed mood (e.g., feels sad, empty, hopeless, tearful).	Y	*Item 2*. Feeling down, depressed, or hopeless.
Markedly diminished interest or pleasure.	Y	*Item 1*. Little interest or pleasure in doing things.
Weight loss when not dieting or weight gain.	N	*Item 5*. Poor appetite or overeating.
Insomnia or hypersomnia.	N	*Item 3*. Trouble falling or staying asleep, or sleeping too much.
Psychomotor agitation or retardation (observable by others).	N	*Item 8*. Moving or speaking so slowly that other people could have noticed? Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual.
Fatigue or loss of energy.	N	*Item 4*. Feeling tired or having little energy.
"Feeling worthless or excessive, delusional or inappropriate guilt.	N	*Item 6*. Feeling bad about yourself—or that you are a failure or have let yourself or your family down.
Fogginess, being unfocused, or indecisive.	N	*Item 7*. Trouble concentrating on things, such as reading the newspaper or watching television.
Thought about harming yourself or suicide.	N	*Item 9*. Thoughts that you would be better off dead or of hurting yourself in some way.

PHQ-9: Patient Health Questionnaire 9 (Kroenke et al., 2001).
DSM-5: Diagnostic and Statistical Manual of Mental Disorders.
Primary symptoms are marked with Y.

TABLE 2 | DSM-5 diagnostic criteria for Generalized Anxiety Disorder.

DSM-5 criteria	Primary	GAD-7 items	PSWQ-8 items
Excessive anxiety and worry, for at least 6 months, about a number of events or activities.	Y	*Item 3*. Worrying too much about different things.	*Item 2*. Many situations make me worry. *Item 4*. When I am under pressure, I worry a lot.
Difficulties in controlling the worry.	Y	*Item 2*. Not being able to stop or control worrying.	*Item 1*. My worries overwhelm me *Item 3*. I know I should not worry about things, but I just cannot help it. *Item 5*. I am always worrying about something. *Item 6*. As soon as I finish one task, I start to worry about everything else I must do. *Item 7*. I have been a worrier all my life *Item 8*. I have been worrying about different things.
Three (or more) of the following six symptoms:			
(1) Restlessness, feeling keyed up or on edge.	N	*Item 1*. Feeling nervous, anxious, or on edge. *Item 5*. Being so restless that it's hard to sit still	
(2) Being easily fatigued.	N	Not represented by any item.	
(3) Difficulty concentrating or mind going blank.	N	Not represented by any item.	
(4) Irritability.	N	*Item 6*. Becoming easily annoyed or irritable.	
(5) Muscle tension.	N	*Item 4*. Trouble relaxing.	
(6) Sleep disturbance.	N		
- *Felling afraid* is not included in the criteria for Generalized Anxiety Disorder 300.02 (F41.1)	N	*Item 7*. Feeling afraid as if something awful might happen.	

Primary symptoms are marked with Y. GAD-7, General Anxiety Disorder-7 (Spitzer et al., 2006); PSWQ-8, Penn State Worry Questionnaire-8 (Hopko et al., 2003).

as for MDD, it can be argued that individuals may focus on the primary, rather than the secondary criteria, when answering the broad word-response questions.

The GAD-7 is developed to capture the DSM-5 criteria for GAD, although the scale does not include items for all symptoms. In addition, it also includes items that are not part of the DSM-5 criteria (for details, see **Table 2**). In contrast, the Penn State Worry Questionnaire-Abbreviated, (PSWQ-8; Hopko et al., 2003) focuses more on worry than the GAD-7. All eight items in the PSWQ-8 include the construct worry (i.e., *worries*, *worry*, *worrying*, or *worrier*; see **Table 2**), whereas GAD-7 comprises two items with the word *worrying* and one with *anxious*. This abbreviated version of the original PSWQ (Meyer et al., 1990; Hopko et al., 2003) is a frequently used measure of worry without the reversed coded items. The PSWQ-8 assesses pathological worry with comparable validity and reliability as the full 16-item version (Wuthrich et al., 2014). As for depression, the word-response question for worry captures individuals' subjective experiences and their own understanding of the construct.

Aims and Hypotheses

This study extends research by Kjell et al. (2019) in two central ways. First, it aims to examine to what extent the QCLA method captures aspects related to the primary and secondary symptoms

associated with MDD and GAD as captured by the items of the corresponding rating scale. To test this, we mapped the word responses to the individual items in the rating scales. The choice of rating scales was motivated because these scales are designed to target the DSM criteria for MDD and GAD (American Psychological Association [APA], 2013). Second, with two rating scales targeting anxiety, the GAD-7 (Spitzer et al., 2006) and the PSWQ-8 (Hopko et al., 2003), we further examined QCLA's ability to capture primary symptoms associated with GAD. These aims are divided into the following three hypotheses.

The Semantic Hypothesis

To further understand the relationship between word responses and rating scales, we examined the correlations to individual items using semantic similarity scales (unipolar and bipolar), language-trained scales, and language-predicted valence scales. We hypothesize that the word embeddings of the word answers significantly capture all items of the depression and worry rating scales through semantic similarity scales and the language-trained scales.

The Valence Hypothesis

Given that the word embeddings for worry and depression words predict the items in the rating scales for the corresponding construct, we further investigated what specific information in the word embeddings contribute to the correlation. Kjell et al. (2019) argued that rating scales are highly influenced by valence, potentially capturing a more general negative feeling for depression and anxiety. Thus, we hypothesize that language-predicted valence scores are correlated with each individual item and that they can explain a substantial part of the correlation between language-trained scales and observed scales.

Primary Over Secondary Criteria Hypothesis

Because the word-response questions focus on respondents' experiences and understanding of a construct, it is important to examine to what extent they also capture the secondary symptoms of the diagnostic criteria for MDD as measured by the PHQ-9 and the GAD as measured by the GAD-7. Based on the relatively general nature of the word-response question (i.e., it does not ask for related symptoms/behaviors of MDD/GAD), we hypothesize that the word embeddings from depression and worry word responses yield stronger correlations to items capturing the primary over the secondary criteria. For example, we anticipate that the semantic similarity scales of depression correlate stronger to the PHQ-9 item about feeling down or depressed (Item 2) than the item about psychomotor agitation or retardation (Item 8).

MATERIALS AND METHODS

Participants

Mechanical Turk (MTurk[1]) was used to recruit participants. This platform enables participants to perform tasks, such as research studies, with an economic gain. MTurk has been used to study clinically relevant topics (Shapiro et al., 2013), with a prevalence of depression and anxiety corresponding to that of community samples [Shapiro et al., 2013; however, other studies suggest higher (Arditte et al., 2016) or lower levels (Veilleux et al., 2015)]. MTurk is generally more diverse than convenience samples such as student and community samples (Chandler and Shapiro, 2016). In our study, 455 respondents submitted their survey, and 44 (9.7%) were excluded from the analyses due to failure to answer the control items correctly (see section "Measures and Material" below). The final sample comprised 411 respondents (47% females, 53% males) ranging in age from 18 to 74 years (Mean = 36.2, *SD* = 11.2). Most participants were from the US (86%), followed by India (11%) and other countries (3%). Out of the 411 participants, 37% were above the cut-off point for MDD on the PHQ-9 (i.e., a score of 10 or higher), and 33% scored above the cut-off for GAD on the GAD-7 (i.e., a score of 10 or higher). These rates were higher than in the general population (Bromet et al., 2011). The participants' reported average perceived personal financial situation was 4.57 (*SD* = 1.71) on a scale ranging from 1 = "Our income does not cover our needs, there are great difficulties" to 7 = "Our income covers our needs, we can save." Participants were paid USD 1 to participate.

Measures and Material

The Word-Response Question of Depression (Kjell et al., 2019) involves asking *Over the last 2 weeks, have you been depressed or not?* coupled with the instructions to answer with their own descriptive words. The instructions furthermore asked participants to "weigh the strength and the number of words" to describe their worry, to focus on writing important and meaningful aspects, and to only write one word in each of the five empty response boxes. Participants were asked to generate five descriptive words.

The Word-Response Question of Worry (Kjell et al., 2019) is coupled with an adapted version of the instructions for the word-response question of depression by changing *depression* to *worry*. i.e., *Over the last 2 weeks, have you been worried or not?* and required five descriptive words as the response format.

The PHQ-9 (Kroenke et al., 2001) includes nine items such as *Feeling down, depressed, or hopeless* coupled with a closed-ended response format ranging from 0 = *Not at all* to 3 = *Nearly every day*. Participants are asked to consider the last 2 weeks. The PHQ-9 has been validated in primary care (Kroenke et al., 2001) and in the general population (Löwe et al., 2004; Martin et al., 2006; Stochl et al., 2020). Additionally, the PHQ-9 has demonstrated the ability to detect changes in response to treatment of various depressive disorders (Löwe et al., 2004). The scale demonstrated a McDonald's ω of 0.94 and a Cronbach's α of 0.92 in the current study.

The GAD-7 (Spitzer et al., 2006) includes seven items such as *Worrying too much about different things* coupled with the same closed-ended response format and timeframe as the PHQ-9. The GAD-7 has been validated in primary care settings (Spitzer et al., 2006) and in the general population (Löwe et al., 2008). Additionally, the GAD-7 has shown sensitivity to detect changes in patients receiving treatment for GAD (Dear et al., 2011). The

[1]www.mturk.com

scale demonstrated a McDonald's ω of 0.95 and a Cronbach's α of 0.93 in the current study.

The PSWQ-8 (Hopko et al., 2003) is an abbreviated version of the PSWQ (Meyer et al., 1990) and encompasses items such as *My worries overwhelm me*, with a closed-ended scale ranging from 0 = *Not at all typical of me* to 5 = *Very typical of me*. In contrast to the full 16-item version, the PSWQ-8 does not include any reverse-coded items. The PSWQ-8 has been validated in a sample of younger (Crittendon and Hopko, 2006) and older (Hopko et al., 2003) adults and a clinical sample of adults (Kertz et al., 2014). The scale yielded a McDonald's ω of 0.97 and a Cronbach's α of 0.96 in the current study.

Control items were randomly presented within the PHQ-9 and the GAD-7, including *On this question please answer the alternative 'Several days,' On this question please answer the alternative 'More than half the days,'* and *On this question please answer the alternative 'Not at all.'* Respondents who failed to answer *all* control items (in total two per participant) correctly were excluded from the analyses. Importantly, attention control items have been found to increase the statistical reliability of the data (e.g., see Oppenheimer et al., 2009; for the use of similar control items see Kjell et al., 2021a).

The demographic survey included questions regarding age, gender, country of origin, first language, and their perception of their household income. When asked about gender, participants were offered three alternatives: male, female, and other. Perceived financial situation was measured by asking, "Does the total income of your household allow you to cover your needs?" with the responses ranging from 1 = "Our income does not cover our needs, there are great difficulties," to 7 = "Our income covers our needs, we can save."

The Word Norm for Depression (Kjell et al., 2019) includes 1,172 words describing being depressed generated by asking 110 participants to describe their "view of being depressed" with 10 words. When constructing the word norm, the targeted word "depression" was also added, so it is the most frequently occurring word by 1 in the norm.

The Word Norm for Worry (Kjell et al., 2019) includes 1,036 words describing being depressed generated by asking 104 participants to describe their "view of being worried" with 10 words. When constructing the norm, the targeted word "worry" was added, so it is the most frequently occurring word by 1 in the norm.

The Word Norm for Not at all Depressed (Kjell et al., 2020) includes 1,125 words generated by 115 participants describing their "view of being not at all depressed."

The Word Norm for Not at all Worried (Kjell et al., 2020) includes 938 words generated by 97 participants describing their "view of being not at all worried."

The language-predicted valence scores were based on a model constructed from the ANEW (Bradley and Lang, 1999), which is a list of more than 1,000 words such as "cat" and "kindness" coupled with participant-rated valence scores ranging from unpleasant to pleasant. The model was created by training the word embeddings for the words in the ANEW list to their corresponding valence score, e.g., cat (M = 4.38, SD = 2.24). Using cross-validation leave-k-out (described below in more detail) produced a strong correlation between predicted and actual valence ratings (r = 0.74, $p < 0.001$, N = 1031 words). This computational model was applied to the word embeddings from the word-response questions in this study to estimate a language-predicted valence score.

Procedure

Participants were informed that the study comprised questions regarding their mental health, including aspects such as depression and worry/anxiety and that they should answer with both descriptive words and rating scales. They were presented with the consent form that included details about how to receive more information, that their responses were recorded anonymously, and that they had the right to withdraw from the study at any time. The study started with the word-response questions presented in random order; followed by the corresponding rating scales in random order. The word-response questions were presented first to avoid the wordings of the rating-scale items from influencing the word responses. Lastly, the participants were asked to fill out the brief demographic survey. In the end, the participants were debriefed. The completion time was on average 10 min and 5 s. According to Swedish law, the National Ethics Committee (protocol number 2020-00730) reviewed the study and decided that it did not require ethical approval.

Statistical Analyses

Natural Language Processing and Machine Learning

The QCLA approach encompasses various techniques to analyze word and text responses in relation to numeric rating scales (see Kjell et al., 2019). These techniques include natural language processing, machine learning, and statistics.

Word embeddings

To represent words with numbers, we used a semantic space from Semantic Excel[2] (Sikström et al., 2018). This space, referred to as *English 1*, was created using an approach akin to latent semantic analysis, where a word co-occurrence table is generated, and the semantic space is produced by applying a data compression algorithm (i.e., SVD) on this table. Technical details on this can be found in Kjell et al. (2019), but see also Landauer and Dumais (1997). This space is generated from the English corpus Google 5-gram database consisting of $1.7 * 10^9$ words (Version 20120701[3]). The generated space consists of the 120,000 most common words in the corpus where each word is represented by a vector consisting of 512 numbers describing how the words in the semantic space are semantically related to each other. This representation is referred to as a *word embedding*.

Responses

Participants' responses were cleaned by changing the word spellings to American English using MS Word, and misspelled words were corrected in those cases where the meaning was clear. The word embeddings for the five words generated by each participant for a given word-response question were aggregated

[2]www.SemanticExcel.com

[3]https://books.google.com/ngrams

by taking the mean across the dimensions, so these words are represented by one word embedding that captures the meaning of the five words taken together in 512 dimensions.

Semantic similarity

A word embedding describes how a word, or set of words, in the semantic space is positioned in relation to all the other words. The closer two words are positioned, the more semantically similar they are. The semantic similarity between two words (or two sets of words) is computed as the cosine of the angle between the two word embeddings in the semantic space. The semantic similarity scores are mathematically bounded between –1 and +1, but in practice they tend to range between a value around 0 (for unrelated words) and a value significantly less than 1, where a higher value indicates higher semantic similarity.

Language training and prediction

The dimensions of the word embeddings may be used as predictors in a multiple regression to predict a numeric variable such as a rating scale. In multiple regression (i.e.,$y = \beta_0 + \beta_1^* x_1 \ldots \beta_m^* x_m + \varepsilon$), y is the observed variable (e.g., the rating scale score), x is the word embedding including several dimensions (i.e., x_1,x_2, etc.), β_m is the coefficient, β_0 the intercept/constant, and ε the error term. For the machine learning implementation, we used the default settings in the *text*-package (version 0.9.10 from CRAN; Kjell et al., 2021b), which involves using ridge regression (Hoerl and Kennard, 1970) with a penalty search grid ranging from 10^{-16} to 10^{16} and a sequence of times 10. The penalty hyperparameter was tuned using 10-fold cross-validation, where the training set was further divided into an analysis (75%) and assessment (25%) set (see Kjell et al., 2021b). This cross-validation procedure enables a determination of the accuracy of the prediction. Here the predicted value ($\hat{y}$) is correlated with the empirical value (y, i.e. the rating scale score), and the correlation coefficient (r) is the measure of accuracy.

Supervised dimension projection plot

The supervised dimension projection (SDP) plot from the *text*-package was used to visualize the word responses. The SDP plots words according to a dimension created by comparing two groups of words (e.g., depression versus worry responses or a quartile split on low versus high scorers on the PHQ-9). In short, the dimension is created by first aggregating all the word embeddings of all words in each group and then subtracting the two aggregated word embeddings to create the *aggregated direction embedding* that is seen to make up a line (dimension) running through origo. Subsequently, individual words' embeddings are first positioned in relation to the mean word embedding of all words (i.e., their word embeddings are subtracted with this embedding) and then projected onto the dimension using the dot product. The *p*-value for each dot product score is computed using a permutation procedure (the default settings in *text* versions.9.10 were used; for more details see Kjell et al., 2021b).

Statistical Software and Packages

The analyses were carried out in R (R Core Team, 2020), where the word-related analyses were carried out using the *text* package (Kjell et al., 2020). Other analyses included using *tidyverse* (Wickham et al., 2019), *psych* (Revelle, 2017), and *Hmisc* (Harrell, 2017).

Interpreting Statistics: Statistical Cut-Off Points

To interpret the internal reliability of the rating scales as good, we used 0.70 as the cut-off for Cronbach's α and McDonald's ω. To interpret the correlation strengths, we used Cohen's (1988) conventions of 0.10, 0.30, and 0.50 for a small/weak, moderate, and large/strong correlation, respectively. Alpha was set to 0.05. The sample size was based on the finding by Kjell et al. (2019) that between 256 and 477 yields correlations to aggregated rating scales that correlate above 0.50 (i.e., $r > 0.5$), which is sufficiently high for evaluating the hypotheses in this article (i.e., the aim here was not to maximize the accuracy but rather to understand the models).

RESULTS

Descriptive Statistics

Descriptive statistics are presented in **Table 3**. The GAD-7 and the PHQ-9 yielded a positive skew and deviated from a normal distribution, whereas the PSWQ-8 demonstrated a normal distribution. Therefore, Spearman rho was applied for the GAD-7 and the PHQ-9 and Pearson's *r* was applied for the PSWQ-8. The correlations among the total scores of the included measures are presented in **Table 4**. **Figures 1A,B** show participants' word responses using SDP plots.

Item Level Analyses of Depression and the PHQ-9

The Semantic Hypothesis

All correlations between the depression bipolar scale and the individual items composing the PHQ-9 were significant and

TABLE 3 | Mean, standard deviation, range, skew, and kurtosis for rating and semantic scales.

Measure	*M*	*SD*	Range	Skew	Kurtosis
PHQ-9	8.41	6.96	27.00	0.51	−0.74
GAD-7	7.37	5.84	21.00	0.33	−0.97
PSWQ-8	25.01	9.87	32.00	−0.30	−1.07
SSS Worry bipolar	0.06	0.28	1.14	−0.67	−0.59
SSS Depression bipolar	−0.01	0.24	0.95	−0.24	−1.22
SSS Depression unipolar	0.28	0.16	0.71	0.51	−0.57
SSS Worry unipolar	0.33	0.19	0.83	0.33	−0.88
Language-trained PHQ-9	8.39	3.85	15.31	−0.49	−1.03
Language-trained GAD-7	7.36	3.04	15.13	−1.03	0.29
Language-trained PSWQ-8	25.03	5.91	24.86	−1.13	0.13
Predicted valence of dep. words	5.21	1.44	5.78	0.19	−1.16
Predicted valence of wor. words	5.04	1.28	5.82	0.26	−1.09

PHQ-9, Patient Health Questionnaire-9; GAD-7, General Anxiety Disorder Scale-7; PSWQ-8, Penn State Worry Questionnaire Abbreviated; SSS, The Semantic Similarity Scale; Bipolar, the semantic similarity of the high norm minus the semantic similarity of the low norm.; Unipolar, the semantic similarity between semantic responses and a high (or low) word norm.; dep., depression; wor., worry.

TABLE 4 | Correlations among measures.

Variable	1	2	3	4	5	6	7	8	9
(1) PHQ-9*									
(2) GAD-7*	0.86***								
(3) PSWQ-8	0.69***	0.81***							
(4) Dw: Bipolar	0.60***	0.53***	0.51***						
(5) Ww: Bipolar	0.44***	0.50***	0.54***	0.55***					
(6) Dw: Unipolar H	0.24***	0.22***	0.23***	0.64***	0.40***				
(7) Ww: Unipolar H	0.23***	0.28***	0.31***	0.38***	0.81***	0.49***			
(8) Dw: Valence	−0.57***	−0.49***	−0.45***	−0.87***	−0.51***	−0.58***	−0.37***		
(9) Ww: Valence	−0.47***	−0.50***	−0.49***	−0.56***	−0.81***	−0.36***	−0.63***	0.52***	

****Indicates $p < 0.001$.*
**For PHQ-9 and GAD-7 we used Spearman's rho.*
PHQ-9, Patient Health Questionnaire-9; GAD-7, General Anxiety Disorder Scale-7; PSWQ-8, Penn State Worry Questionnaire Abbreviated; Dw, depression words; Ww, worry words; Bipolar, bipolar semantic similarity scale; Unipolar, unipolar semantic similarity scale; H, High; Valence, language-predicted scale ANEW valence.
Rows 1–7 of this correlation table are also presented in Kjell et al. (2020).

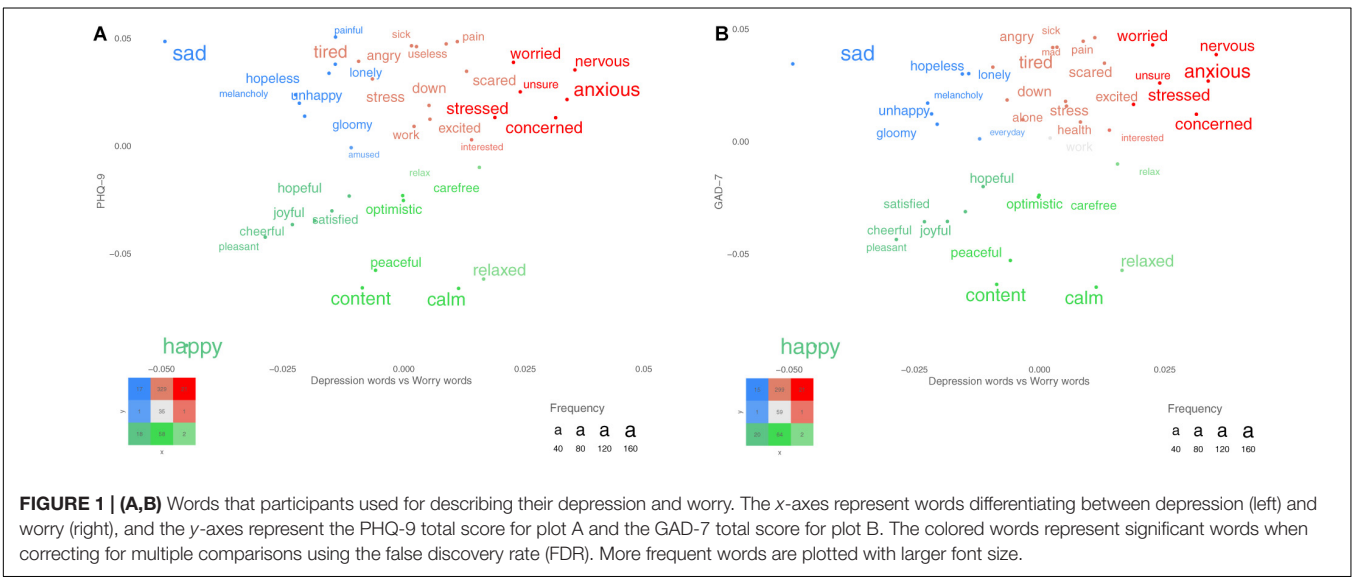


FIGURE 1 | (A,B) Words that participants used for describing their depression and worry. The *x*-axes represent words differentiating between depression (left) and worry (right), and the *y*-axes represent the PHQ-9 total score for plot A and the GAD-7 total score for plot B. The colored words represent significant words when correcting for multiple comparisons using the false discovery rate (FDR). More frequent words are plotted with larger font size.

varied from moderate to strong in correlational strengths (**Table 5**; bipolar:ρ = 0.31–0.61). In comparison to the bipolar scale, the unipolar scale showed lower correlations to individual items (ρ = 0.11–0.31), with non-significant correlations for Item 8 (moving patterns) and Item 9 (self-harm).

The results for language-trained PHQ scales were similar to the findings for semantic similarity scales. Training the responses for the word-response question of depression to the individual items composing the PHQ-9 yielded significant correlations ranging from weak to strong (ρ = 0.30–0.60; $p < 0.001$; see **Table 6**).

The Valence Hypothesis

The word responses' language-predicted valence scale also correlated significantly with each PHQ-9 item (ρ = –0.30 to –0.57, $p < 0.001$). Absolute values for individual items except for Items 3 and 8 were stronger for the language-trained scales compared with the language-predicted valence scale. Controlling for valence reduced the language-trained item level correlations substantially from mean ρ = 0.47 (range:ρ = 0.30–0.60) to mean ρ = 0.22 (range: ρ = 0.12–0.29, see **Table 6**).

The Primary Over Secondary Criteria Hypothesis

The highest correlation for both the unipolar and the bipolar scale was to the item tapping into feeling down and depressed (Item 2). The lowest correlation for the unipolar scale was to the item targeting concentration difficulties (Item 7). The lowest corelations for the bipolar scale were to moving patterns (Item 8) and self-harm (Item 9). Changes in moving pattern (Item 8) and thoughts about being better off dead and self-harm (Item 9) were significant for the bipolar semantic similarity scale but not for the unipolar semantic similarity scale. The strongest correlations for both language-trained scales and language-predicted valence scales were to an item tapping into the emotional experience of depression (i.e., feeling down; ρ = 0.60 and –0.57, respectively). The lower correlations for both language-trained scales and language-predicted valence scales tended to be to behaviors relating to

TABLE 5 | Spearman's rho correlations between total and individual PHQ-9 items for the unipolar (high) and bipolar semantic similarity scales for depression.

PHQ-9	*M*	*SD*	Item total correlation	Unipolar SSS depression	Bipolar SSS depression
Total	8.41	6.69		0.24***	0.60***
Item 1 (little interest)	0.95	0.96	0.77***	0.20***	0.53***
Item 2 (feeling down)	1.01	0.99	0.80***	0.27***	0.61***
Item 3 (disrupted sleep)	1.16	1.02	0.67***	0.26***	0.49***
Item 4 (little energy)	1.26	0.99	0.72***	0.31***	0.55***
Item 5 (changed food habits)	0.91	1.00	0.74***	0.23***	0.50***
Item 6 (failure)	1.03	1.03	0.78***	0.16**	0.50***
Item 7 (no concentration)	0.93	1.02	0.77***	0.11*	0.44***
Item 8 (slow or restless)	0.58	0.87	0.62***	0.08	0.31***
Item 9 (self-harm)	0.58	0.93	0.68***	0.5	31***

N = 411.
**Indicates p < 0.05; **indicates p < 0.01; ***indicates p < 0.001.*
M and SD are the mean and standard deviation, respectively.
PHQ-9, Patient Health Questionaire-9; SSS, Semantic Similarity Scale.

TABLE 6 | The mean and standard deviation for the PHQ-9 and its individual items and their correlations to the language-trained scale and the language-predicted valence scale from the depression word responses.

PHQ-9 items	*M*	*SD*	Language-trained PHQ	Language-predicted valence	Partial correlation controlling for valence
PHQ-9 total	8.41	6.96	0.60***	−0.57***	0.29***
Item 1 (little interest)	0.95	0.96	0.54***	−0.50***	0.29***
Item 2 (feeling down)	1.01	0.99	0.60***	−0.57***	0.28***
Item 3 (disturbed sleep)	1.16	1.02	0.44***	−0.46***	0.12*
Item 4 (little energy)	1.26	0.99	0.51***	−0.50***	0.21***
Item 5 (changed food habits)	0.91	1.00	0.47***	−0.44***	0.23***
Item 6 (failure)	1.03	1.03	0.50***	−0.49***	0.23***
Item 7 (no concentration)	0.93	1.02	0.45***	−0.40***	0.26***
Item 8 (slow or restless)	0.58	0.87	0.30***	−0.30***	0.15**
Item 9 (self-harm)	0.58	0.93	0.38***	−0.34***	0.22***

N = 411.
**Indicates p < 0.05; **indicates p < 0.01; ***indicates p < 0.001.*
Spearman's rho, using Holms correction for multiple comparison. M and SD are the mean and standard deviation, respectively; PHQ-9, Patient Health Questionnaire-9, with item numbers corresponding to the order in Kroenke et al. (2001). Partial correlation controlling for valence = Partialling out the language-predicted valance between the correlation of the language-trained PHQ score and the observed PHQ score.

depression; for example, see Item 8 (moving pattern;ρ = 0.30 and −0.30, respectively) and Item 9 (self-harm;ρ = 0.38 and −0.34, respectively).

TABLE 7 | Spearman's rho correlations between total and individual GAD-7 items and the unipolar (high) and bipolar semantic similarity scales for worry.

GAD-7	*M*	*SD*	Item total correlation for GAD-7	Unipolar SSS Worry	Bipolar SSS Worry
Total	7.37	5.84	–	0.28***	0.50***
Item 1 (anxious, on edge)	1.13	0.97	0.88***	0.32***	0.50***
Item 2 (cannot control worrying)	1.09	1.03	0.90***	0.26***	0.47***
Item 3 (worrying about different things)	1.15	1.00	0.90***	0.27***	0.47***
Item 4 (trouble relaxing)	1.14	1.00	0.86***	0.21***	0.41***
Item 5 (restlessness, hard to sit still)	0.84	0.96	0.79***	0.16**	0.32***
Item 6 (annoyed, irritable)	1.05	0.94	0.78***	0.24***	0.42***
Item 7 (afraid)	0.97	0.99	0.85***	0.21***	0.41***

N = 411.
***Indicates p < 0.01; ***indicates p < 0.001.*
M and SD are the mean and standard deviation, respectively.
GAD-7, Generalised Anxiety Disorder Scale-7; SSS, Semantic Similarity Scale; item numbers correspond to the order in Spitzer et al. (2006).

Item Level Analyses of Worry and the GAD-7 and the PSWQ-8

The Semantic Hypothesis

All correlations of the worry unipolar and bipolar scales to each of the individual items composing the GAD-7 were significant and ranged from weak to strong (**Table 7**; unipolar:ρ = 0.16–0.32; bipolar:ρ = 0.32–0.50). All items composing the PSWQ-8 also correlated significantly with the worry unipolar and bipolar scales (**Table 8**; unipolar r = 0.26–0.34; bipolar: r = 0.52–0.61).

The results from training the worry word responses to the individual items of the GAD-7 showed overall moderate correlations ranging from ρ = 0.41 to **0.52** ($p < 0.001$; see **Table 9**), and training to the items composing the PSWQ-8 resulted in moderate correlations ranging from r = 0.52 to **0.63** ($p < 0.001$; see **Table 10**).

The Valence Hypothesis

The language-predicted valence scale from the worry responses also showed significant correlations to each of the individual items of the GAD-7 (r = −0.31 to −0.50) and the PSWQ-8 items (r = −0.42 to r = −0.55, $p < 0.001$). Controlling for valence using partial correlation reduced the language-trained item level correlations substantially for GAD-7 [mean ρ = 0.46 (range ρ = 0.41–0.52) to meanρ = 0.23 (range ρ = 0.19 −0.31)] and for PSWQ-8 [mean r = 0.58 (range r = 0.52–0.59) to mean r = 0.38 (range r = 0.37–0.40)].

Primary Over Secondary Criteria Hypothesis

The highest correlations for both the unipolar and the bipolar scales were to the GAD-7 item representing the primary criterion

TABLE 8 | Pearson correlations between total and individual PSWQ-8 items the unipolar (high) and bipolar semantic similarity scales for worry.

PSWQ-8 items	*M*	*SD*	Item total correlation	Unipolar SSS Worry	Bipolar SSS Worry
Total	25.01	9.87		0.33***	0.6***
Item 1 (worries overwhelming)	2.96	1.39	0.87***	0.27***	0.58***
Item 2 (situations worry)	3.15	1.38	0.91***	0.28***	0.55***
Item 3 (worry different things)	3.24	1.41	0.89***	0.30***	0.60***
Item 4 (pressure, worry)	3.39	1.32	0.86***	0.31***	0.55***
Item 5 (always worrying)	3.03	1.43	0.93***	0.30***	0.57***
Item 6 (after task, worry about new)	2.96	1.36	0.85***	0.28***	0.54***
Item 7 (worrier all my life)	3.08	1.47	0.85***	0.26***	0.52***
Item 8 (worrying about things)	3.21	1.38	0.87***	0.33***	0.53***

N = 411.
****Indicates p < 0.001.*
M and SD are the mean and standard deviation, respectively.
PSWQ-8, Penn State Worry Questionnaire Abbreviated; SSS, Semantic Similarity Scale; item numbers correspond to the order in Crittendon and Hopko (2006).

TABLE 9 | The mean and standard deviation for the GAD-7 and its individual items and their correlations to the language-trained scale and the language-predicted valence scale from the worry word responses.

GAD-7 items	*M*	*SD*	Language-trained GAD	Language-predicted valence	Partial correlation controlling for valence
GAD-7 total	7.37	5.84	0.54***	−0.50***	0.29***
Item 1 (anxious, on edge)	1.13	0.97	0.49***	−0.47***	0.24***
Item 2 (excessive worrying)	1.09	1.03	0.52***	−0.49***	0.25***
Item 3 (different areas of worry)	1.15	1.00	0.50***	−0.50***	0.19***
Item 4 (trouble relaxing)	1.14	1.00	0.41***	−0.39***	0.21***
Item 5 (restlessness)	0.84	0.96	0.42***	−0.31***	0.31***
Item 6 (annoyed, irritable)	1.05	0.94	0.42***	−0.40***	0.20***
Item 7 (afraid)	0.97	0.99	0.45***	−0.44***	0.23***

N = 411.
****Indicates p < 0.001.*
Spearman's rho, using Holms correction for multiple comparison. M and SD are the mean and standard deviation, respectively.
Ww, worry words, GAD-7, Generalised Anxiety Disorder Scale-7, item numbers correspond to the order in Spitzer et al. (2006). Partial correlation controlling for valence = Partialling out the language-predicted valance between the correlation of language-trained GAD-7 score and observed GAD score.

about *feeling anxious and on edge* (Item 1) followed by the items about *worrying too much about different things* (Item 3) and worry

TABLE 10 | The mean and standard deviation for the PSWQ-8 and its individual items and their correlations to the language-trained scale and the language-predicted valence scale from the worry word responses.

PSWQ-8 items	*M*	*SD*	Language-trained PSWQ-8	Language-predicted valence	Partial correlation controlling for valence
PSWQ-8 total	25.01	0.87	0.66***	−0.54***	0.46***
Item 1 (worries are overwhelming)	0.97	0.97	0.59***	−0.49***	0.39***
Item 2 (worrying about situations)	3.15	1.38	0.56***	−0.45***	0.39***
Item 3 (worrying about different things)	3.24	1.41	0.61***	−0.51***	0.40***
Item 4 (pressure, worry)	3.39	1.32	0.56***	−0.46***	0.37***
Item 5 (always worrying)	3.03	1.43	0.59***	−0.49***	0.39***
Item 6 (after completing a task, worrying about the next one)	2.96	1.36	0.56***	−0.45***	0.38***
Item 7 (worrier all my life)	3.08	1.47	0.52***	−0.42***	0.34***
Item 8 (worrying about things)	3.21	1.38	0.63***	−0.55***	0.39***

N = 411.
****Indicates p < 0.001.*
Pearson correlation, using Holms correction for multiple comparison. M and SD are the mean and standard deviation, respectively.
Ww, worry words; PSWQ-8, Penn State Worry Questionnaire Abbreviated, with item numbers corresponding to the order in Crittendon and Hopko (2006); Partial correlation controlling for valence = Partialling out the language predicted valance between the correlation of the language-trained PSWQ score and the observed PSWQ score.

not being able to stop or control worrying (Item 2). The lowest correlation was to the GAD-7 item tapping into the secondary criterion regarding *difficulties sitting still* (Item 5; i.e., a behavior).

For the PSWQ-8, all eight items demonstrated similar strengths (which were comparable in magnitude to the strongest items in the GAD-7). This consistency in strength and the comparably strong correlation make sense considering that all items are quite similar, tapping into the primary criterion of excessive worrying with different forms of the word "worry." The highest correlation for both the unipolar and bipolar worry scale was in relation to Item 8, which targets the general tendency to worry.

In terms of the language-trained GAD-7 and language-predicted valence scales, the strongest correlations were to the item tapping into experiencing worry in different areas in life (ρ = 0.50 and –0.50 for the language-trained scales and the language-predicted valence scale, respectively). The lowest correlation was to items representing the secondary criteria about trouble relaxing (Item 4, ρ = 0.41 and –0.39) and being so restless that one finds it difficult to sit still (Item 5, ρ = 0.42 and –0.31), which can be seen as a related behavior to the experience of worry.

DISCUSSION

The Semantic Hypothesis

We examined the relationship between word responses and rating scales by correlating the individual items of the respective rating scales with language-trained scales and semantic similarity scales (unipolar and bipolar). The semantic hypothesis was supported for the language-trained scales and the bipolar semantic similarity scales for the PHQ-9, the GAD-7, and the PSWQ-8. The unipolar semantic similarity scales correlated significantly for the GAD-7 and the PSWQ-8 items, but not for all items of the PHQ-9. Overall, these findings suggest that word embeddings capture the diagnostics criteria for MDD and GAD as measured by two of three QCLAs.

The Valence Hypothesis

In accordance with the valence hypothesis, the language-predicted valence scales correlated significantly with each of the items composing the PHQ-9, the GAD-7, and the PSWQ-8. The language predicted valence scale overall tended to show comparable correlations to all items comprising the PHQ-9, equal or lower correlations to the GAD-7 and the PSWQ-8 items as compared with the language trained scales. A large part of the language-trained correlation was accounted for by the language-predicted valence scales; however, a significant portion of the correlation remained for all items. The results suggest that valence is a potentially strong contributor carrying a large part of the information for predicting rating scale items.

The Primary Over Secondary Criteria Hypothesis

The hypothesis that the QCLA word-response would capture the primary criteria – i.e., the subjective experiences (such as thoughts and feelings) – better than secondary criteria (often more behavioral aspects) was supported. For the depression semantic similarity scales, the strongest correlation was to an item about feeling tired and having little energy. For the language-trained scales for depression, the strongest correlation was to the item tapping into feeling depressed, down, and hopeless, whereas the weakest correlations for both the language-trained scale and the semantic similarity scale for depression were to the items about related behavioral symptoms, including having trouble concentrating (Item 7), moving slowly or being fidgety (Item 8), and thoughts about being better off dead or self-harming (Item 9).

Further, it appears that the semantic similarity scales for worry primarily corresponded to the aspects of the rating scales that capture the primary criteria and the subjective experiences and understandings of worry, and less so the secondary criteria (especially the related behaviors). That is, the unipolar and bipolar scales correlated strongest to the GAD-7 items about feeling anxious and worry (Items 1–3) and less strongly to items tapping into related behaviors such as not being able to sit still (Item 5). The semantic similarity scale for worry captures all the items in the PSWQ-8 to a similar degree, which could be explained by all of the items directly asking about *worry*, *worrying*, *worries*, or being a *worrier*.

Overall, these results support the primary over secondary criteria hypothesis that the QCLA correlates stronger with items tapping into the cognitive experience of worry rather than with related behaviors. Thus, the QCLA appears to primarily capture the subjective experience of MDD and GAD, whereas when related behaviors appear to be captured they are covered less well. This is noteworthy from a clinical perspective and suggests that future studies are warranted to test specific word-response questions or word norms to capture behavioral aspects of the DSM-5 criteria for MDD and GAD. These findings suggest that future research should also consider developing word-response questions and/or word norms that more accurately capture secondary symptoms. For example, to measure the related behavior of self-harm, one could develop a word-response question that explicitly asks individuals to describe whether they self-harm or not, or alternatively creating a word norm comprising words related to self-harm and applying it to the word responses of the question on depression.

Potential for Clinical Significance

The QCLA method allows participants to directly express and describe their experiences freely. In addition, the QCLA method does not prime patients with symptoms that the patient does not necessarily have or that are irrelevant for the patient. Thus, QCLA may have the potential to add value to clinical research and practices by complementing traditional rating scales and enhancing our understanding of patients' experiences.

Research suggests that rating scales of depression tend to fail in reliably capturing the disorder across scales and that the scales miss important symptoms. For example, a literature review points out that seven commonly used rating scales for depression do not necessarily measure the same disorder because these scales include items that are aimed to measure a wide range of different symptoms (Van Loo et al., 2012; Fried, 2017). From this it follows that constructs do not generalize across scales that are aimed to measure the same disorder. Furthermore, to identify symptoms related to depression that matter for patients, Chevance et al. (2020) asked participants to describe "the most difficult aspect of depression to live with or endure?". They found that the most frequently mentioned symptom was "mental pain," which is missing in the DSM-5 criteria and in the closed-ended rating scale for depression used in this study. Importantly, the QCLA method offers an opportunity to examine the presence of symptoms beyond primary and secondary criteria (e.g., note that *painful* is one of the descriptive words related to high depression and depression responses in **Figure 1A** and that *pain* is related to both high PHQ-9 and GAD-7 in **Figures 1A,B**).

Semantic Similarity and Valence

In clinical settings, the language-predicted valence scales may potentially complement the semantic similarity scales in important ways. Language-predicted valence scores could be used to signal that further investigations are required even though

the semantic similarity score is low (i.e., outside established cut-off points). That is, when answering a word-response question a person may use words that are comparably distant from the targeted word norm, but that from a mental health perspective warrant further investigation. In the cases where words have a negative valence, the language-predicted valence scale may be used to alert further attention. For example, to the word-response question for worry a patient may answer something that is semantically relatively far away from the worry word norm because it is related to, for example, depression; however, alarming words will often have a negative valence. Hence, if word responses have a low semantic similarity but a high negative valence score, this signals that the respondent's answer should be investigated more closely and that another set of word-response questions may be needed.

Language Models and Objective Measures

We argue that language-trained scales are valuable when investigating the relationship between word-responses and numerical rating scales. In clinical settings, language-trained scales and predictions will potentially be very important given that they can be trained and used to predict objective measures and outcomes rather than self-reported numerical rating scales. That is, word responses may be trained to actual behavior such as sick leaves, number of suicide attempts, or information obtained from smartphone apps including quality of sleep, walking speed, etc. [e.g., see Miller (2012) for ways to collect data with smartphones]. These language-trained models can potentially be used to investigate the relationship between word responses and the objective measure as well as to predict future respondents' behaviors with the possibility to tailor treatment interventions.

Limitations

It is important to note that this study used *individual* items to examine the degree that QCLAs capture symptoms; hence, future research could use specific assessments (rather than just one item) and/or objective measures (e.g., of sleep) to examine this further. Future studies could also examine potential benefits from using recent language models that can take word order (i.e., context) into account (e.g., Devlin et al., 2019), which is an improvement by analyzing descriptive texts as compared to descriptive words (Kjell et al., in progress).

MTurk is an efficient way to collect data from individuals with a wide range of backgrounds. It should be noted that generalization from MTurk should be made carefully; however, previous MTurk studies have been shown to be more representative compared with other samples commonly used in clinical research (e.g., Chandler and Shapiro, 2016). Lastly, this study did not collect data that allows for analysis of attrition (i.e., only data where participants completed the entire survey was collected), which further emphasizes the importance of being careful when generalizing the results.

CONCLUSION

The QCLAs (i.e., unipolar, bipolar, language trained, and language-predicted valence scales) cover all aspects of the rating scale items that are designed to cover the primary DSM criteria as measured by rating scales, although with strengths varying from weak to strong. The QCLAs appear specifically suited to capture an individual's cognitive and emotional experiences of depression, worry, and anxiety. Overall, they also capture the secondary criteria (generally including more behaviors and physiological symptoms) related to these experiences as measured by the rating scales. We believe that the QCLAs could be of great importance for clinical research and practice, where word-responses are coupled with objectively measured outcomes. Further, because the QCLA method is based on the respondent's own descriptions of their experiences and symptoms related to a construct, the method carries the potential to personalize assessments, which might contribute to the ongoing discussion regarding the diagnostic heterogeneity of depression.

DATA AVAILABILITY STATEMENT

Data will not be made freely available as participants where not informed about this, and these open-ended text responses may be a risk for identification.

ETHICS STATEMENT

The study was reviewed by the Ethics Review Authority who decided that it did not require ethical approval. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KK has stated the hypotheses, collected the data, performed the analysis, and written the manuscript with the support and guidance from SS. PJ contributed to help the stating hypothesis and gave feedback to the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by Vinnova Sweden's National Innovation Agency, d.no. 2018-02007.

REFERENCES

American Psychological Association [APA] (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Washington, DC: American Psychiatric Association.

Arditte, K. A., Çek, D., Shaw, A. M., and Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychol. Assessm.* 28:684.

Bradley, M. M., and Lang, P. J. (1999). *Affective Norms for English Words (ANEW): INSTRUCTION MANUAL and Affective Ratings*. Technical Report C-1. Florida: The Center for Research in Psychophysiology, University of Florida.

Bromet, E., Andrade, L. H., Hwang, I., Sampson, N. A., Alonso, J., De Girolamo, G., et al. (2011). Cross-national epidemiology of DSM-IV major depressive episode. *BMC Med.* 9:90. doi: 10.1186/1741-7015-9-90

Chandler, J., and Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annu. Rev. Clin. Psychol.* 12, 53–81. doi: 10.1146/annurev-clinpsy-021815-093623

Chevance, A., Ravaud, P., Tomlinson, A., Le Berre, C., Teufer, B., Touboul, S., et al. (2020). Identifying outcomes for depression that matter to patients, informal caregivers, and health-care professionals: qualitative content analysis of a large international online survey. *Lancet Psychiatry* 7, 692–702.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioralbehavioural Sciences*. Hilsdale, NJ: Lawrence Earlbaum Associates.

Crittendon, J., and Hopko, D. R. (2006). Assessing worry in older and younger adults: psychometric properties of an abbreviated Penn State Worry Questionnaire (PSWQ-A). *J. Anxiety Disord.* 20, 1036–1054.

de Beurs, E., den Hollander-Gijsman, M. E., van Rood, Y. R., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., et al. (2011). Routine outcome monitoring in the Netherlands: practical experiences with a web- based strategy for the assessment of treatment outcome in clinical practice. *Clin. Psychol. Psychother.* 18, 1–12.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). "Predicting depression via social media," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media ICWSM-13*, New York, NY.

Dear, B. F., Titov, N., Sunderland, M., McMillan, D., Anderson, T., Lorian, C., et al. (2011). Psychometric comparison of the generalized anxiety disorder scale-7 and the Penn State Worry Questionnaire for measuring response during treatment of generalised anxiety disorder. *Cogn. Behav. Therapy* 40, 216–227.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1810.04805

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., et al. (2018). Facebook language predicts depression in medical records. *Proc. Natl. Acad. Sci. U.S.A.* 115, 11203–11208.

Fried, E. I. (2017). The 52 symptoms of major depression: lack of content overlap among seven common depression scales. *J. Affect. Disord.* 208, 191–197.

Harrell, F. E. Jr. (2017). *Hmisc: Harrell Miscellaneous. R Package Version 4.0-3*. Available online at: https://CRAN.R-project.org/package=Hmisc

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.

Hopko, D. R., Reas, D. L., Beck, J. G., Stanley, M. A., Wetherell, J. L., Novy, D. M., et al. (2003). Assessing worry in older adults: confirmatory factor analysis of the Penn state worry questionnaire and psychometric properties of an abbreviated model. *Psychol. Assess.* 15:173.

Kertz, S. J., Lee, J., and Björgvinsson, T. (2014). Psychometric properties of abbreviated and ultra-brief versions of the Penn State Worry Questionnaire. *Psychol. Assess.* 26:1146.

Kesebir, P., and Diener, E. (2008). In pursuit of happiness empirical answers to philosophical questions. *Perspect. Psychol. Sci.* 3, 117–125. doi: 10.1111/j.1745-6916.2008.00069.x

Kjell, O. N., and Diener, E. (2020). Abbreviated three-item versions of the satisfaction with life scale and the harmony in life scale yield as strong psychometric properties as the original scales. *J. Pers. Assess.* 103, 183–194.

Kjell, O. N. E., Daukantaitë, D., and Sikstroöm, S. (2021a). Computational language assessments of harmony in life -not satisfaction with life nor rating scales - correlate with cooperative behaviors. *Front. Spec. Issue Seman. Algor. Assess. Attitud. Personal.* doi: 10.3389/fpsyg.2021.601679

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2019). Semantic measures: using natural language processing to measure, differentiate, and describe psychological constructs. *Psychol. Methods* 24, 92–115.

Kjell, O. N. E., Kjell, K., Garcia, D., and Sikström, S. (2020). "Semantic similarity scales: using semantic similarity scales to measure depression and worry," in *Statistical Semantics*, eds S. Sikström and D. Garcia (Cham: Springer), doi: 10.1007/978-3-030-37250-7_4

Kjell, O. N. E., Giorgi, S., and Schwartz, H. A. (2021b). *Text: An R-Package for Analyzing and Visualizing Human Language Using Natural Language Processing and Deep Learning*. Available online at: https://doi.org/10.31234/osf.io/293kt (accessed April 16, 2021).

Kjell, O. N. E., Sikström, S., Kjell, K., and Schwartz, A. (in progress). *Question Based Text Responses Analysed with AI-Based Transformer Approach the Upper Limits of Rating Scale Reliability*.

Kroenke, K., Spitzer, R. L., and Williams, J. W. B. (2001). The PHQ-9: validity of a brief depression severity measure [Electronic version]. *J. Gen. Intern. Med.* 16, 606–613.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295x.104.2.211

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., et al. (2008). Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population. *Med. Care* 46, 266–274.

Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., and Kroenke, K. (2004). Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med. Care* 42, 1194–1201.

Martin, A., Rief, W., Klaiberg, A., and Braehler, E. (2006). Validity of the brief patient health questionnaire mood scale (PHQ-9) in the general population. *Gen. Hosp. Psychiatry* 28, 71–77.

Meyer, T. J., Miller, M. L., Metzger, R. L., and Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behav. Res. Therapy* 28, 487–495.

Miller, G. (2012). The smartphone psychology manifesto. *Perspect. Psychol. Sci.* 7, 221–237.

Mowery, D., Smith, H. A., Cheney, T., Bryan, C., and Conway, M. (2016). Identifying depression-related tweets from twitter for public health monitoring. *Online J. Public Health Inform.* 8:6561. doi: 10.5210/ojphi.v8i1.6561

Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exp. Soc. Psychol.* 45, 867–872.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., Langer, E. J., et al. (2017). Forecasting the onset and course of mental illness with Twitter data. *Sci. Rep.* 7:13006. doi: 10.1038/s41598-017-12961-9

Revelle, W. (2017). *Psych: Procedures for Personality and Psychological Research*. Evanston, IL: Northwestern University.

Shapiro, D. N., Chandler, J., and Mueller, P. A. (2013). Using mechanical turk to study clinical populations. *Clin. Psychol. Sci.* 1, 213–220.

Sikström, S., Kjell, O. N. E., and Kjell, K. (2018). *Semantic Excel: An Introduction to a User-Friendly Online Software Application for Statistical Analyses of Text Data*. Available online at: https://doi.org/10.31234/osf.io/z9chp (accessed October 25, 2018).

Spitzer, R. L., Kroenke, K., Williams, J. W. B., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the gad-7. *Archiv. Intern. Med.* 166, 1092–1097. doi: 10.1001/archinte.166.10.1092

Stochl, J., Fried, E. I, Fritz, J., Croudace, T. J., Russo, D. A., Knight, C., et al. (2020). On dimensionality, measurement invariance, and suitability of sum scores for the PHQ-9 and the GAD-7. *Assessment* doi: 10.1177/1073191120976863

Van Loo, H. M., De Jonge, P., Romeijn, J. W., Kessler, R. C., and Schoevers, R. A. (2012). Data-driven subtypes of major depressive disorder: a systematic review. *BMC Med.* 10:156. doi: 10.1186/1741-7015-10-156

Veilleux, J. C., Salomaa, A. C., Shaver, J. A., Zielinski, M. J., and Pollert, G. A. (2015). Multidimensional assessment of beliefs about emotion: development and validation of the emotion and regulation beliefs scale. *Assessment* 22, 86–100.

Washington, A. E., and Lipstein, S. H. (2011). The patient-centered outcomes research institute–promoting better information, decisions, and health. *N. Engl. J. Med.* 365:e31.

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino, L., McGowan, L., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4:1686.

Wuthrich, V. M., Johnco, C., and Knight, A. (2014). Comparison of the Penn State Worry Questionnaire (PSWQ) and abbreviated version (PSWQ-A) in a clinical and non-clinical population of older adults. *J. Anx. Disord.* 28, 657–663. doi: 10.1016/j.janxdis.2014.07.005

Conflict of Interest: KK and SS co-founded WordDiagnostics focusing on diagnosing psychiatric disorders using language based assessments.

The remaining 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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership