METHODOLOGICAL ISSUES IN PSYCHOLOGY AND SOCIAL SCIENCES RESEARCH

EDITED BY: Begoña Espejo, Marta Martín-Carbonell and Irene Checa
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-83250-578-6
DOI 10.3389/978-2-83250-578-6

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

METHODOLOGICAL ISSUES IN PSYCHOLOGY AND SOCIAL SCIENCES RESEARCH

Topic Editors:
Begoña Espejo, University of Valencia, Spain
Marta Martín-Carbonell, Universidad Cooperativa de Colombia, Colombia
Irene Checa, University of Valencia, Spain

Citation: Espejo, B., Martín-Carbonell, M., Checa, I., eds. (2022). Methodological Issues in Psychology and Social Sciences Research. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-578-6

Table of Contents

frontiers in Psychology

OPINION
published: 23 November 2021
doi: 10.3389/fpsyg.2021.783226

The Ethical Use of Fit Indices in Structural Equation Modeling: Recommendations for Psychologists

*Bryant M. Stone**

Department of Psychology, Southern Illinois University Carbondale, Carbondale, IL, United States

Fit indices provide helpful information for researchers to assess the fit of their structural equation models to their data. However, like many statistics and methods, researchers can misuse fit indices, which suggest the potential for questionable research practices that might arise during the analytic and interpretative processes. In the current paper, the author highlights two critical ethical dilemmas regarding the use of fit indices, which are (1) the selective reporting of fit indices and (2) using fit indices to justify poorly-fitting models. The author highlights the dilemmas and provides potential solutions for researchers and journals to follow to reduce these questionable research practices.

Keywords: structural equation modeling, factor analysis, ethical issues, model fit, fit indices

OPEN ACCESS

Edited by:
Marta Martín-Carbonell, Universidad Cooperativa de Colombia, Colombia

Reviewed by:
Steffen Zitzmann, University of Tübingen, Germany

Correspondence:
Bryant M. Stone
Bryant.Stone@siu.edu

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

Received: *25 September 2021*
Accepted: *19 October 2021*
Published: *23 November 2021*

Citation:
Stone BM (2021) The Ethical Use of Fit Indices in Structural Equation Modeling: Recommendations for Psychologists. Front. Psychol. 12:783226. doi: 10.3389/fpsyg.2021.783226

INTRODUCTION

Structural equation modeling (SEM) allows researchers to analyze data in ways that are impossible under the general linear model, such as simultaneously assessing multiple relationships across variables or measuring variables that researchers cannot directly observe (i.e., latent variables) such as depression or self-esteem. Many modern scales and measures within the social sciences and education, such as intelligence tests, personality assessments, and diagnostic tools for mental health professionals, use structural equation modeling to align measures with underlying latent constructs. Researchers must create a model, collect the data, and then test the model's fit to the collected data. Although there are many ways to assess for model fit, many researchers rely on fit indices, a collection of statistics that quantify the degree of data-model fit. These measures may assist researchers in judging the fit of their models. However, like many statics and methods, researchers may misuse fit indices through unethical and questionable research practices. The current paper investigates and suggests future directions for two ethical dilemmas regarding fit indices: the selective reporting of fit indices to bias the apparent fit of a model and the use of fit indices to justify poorly fitted models.

FIT INDICES

Researchers have categorized the dozens of fit indices into four broad domains (Hu and Bentler, 1999). First, researchers calculate absolute fit indices (e.g., standardized root-mean-square residual) by comparing the observed covariance matrix (i.e., the collected data) to the implied covariance matrix (i.e., the covariances that arose from the specified model). Second, relative fit indices (e.g., Tucker-Lewis Index) compare the specified model to a baseline model.

A baseline model is a model where all the observed variables or collected data are uncorrelated. Third, noncentrality-based indices (e.g., Comparative Fit Index or the Root-Mean-Square Error of Approximation) are indices that adjust the perfect fit of the model, so that the chi-square equals the model's degrees of freedom instead of zero. Fourth, parsimonious fit indices (e.g., Parsimonious Goodness of Fit Index) tend to be fit indices from other categories adjusted to favor more parsimonious models over more complex models. These fit indices quantify the model fit through multiple methods.

THE SELECTIVE REPORTING OF FIT INDICES

Fit indices are easily biased and demonstrate considerable variability. Some fit indices are less vulnerable to the influence of extraneous variables, such as the CFI and RMSEA (Cangur and Ercan, 2015). However, some estimation techniques significantly inflate the standardized root mean squared residual (SRMR; e.g., generalized least squares technique is inflated compared to the asymptotically distribution-free technique). Other studies have found that the sample size easily biases the Tucker-Lewis index (TLI) and the normed fit index (NFI; Yadama and Pandey, 1995). Further, some fit indices, such as the CFI, TLI, and RMSEA, are biased to favor bifactor models (Morgan et al. 2015). The varying sensitivity to extraneous factors increases the amount of variability across fit indices.

The significant variability across fit indices may influence researchers to report those indices that suggest the best model fit. Many SEM software packages (e.g., *R* or LISREL) automatically calculate multiple fit indices when performing the initial SEM analyses. The automatic calculation of multiple fit indices allows researchers to observe and report the fit indices that support their model's fit. For example, individuals with more complicated models may choose not to report parsimonious fit indices, which favor simpler models; and individuals with larger sample sizes may choose to report the NFI or NNFI, which favor models when sample sizes are large. This selective reporting may mislead the readers to believe the specified model fits the data better than it does. Therefore, fit indices provide a wide range of useful information about the data-model fit; however, researchers may engage in questionable research practices by selectively reporting certain fit indices.

FUTURE DIRECTIONS

Two potential solutions may limit researchers' ability to selectively report fit indices that justify their model. First, journals may consider standardizing the fit indices that they publish in their journal. Journals tend to have few standards for publishing SEM analyses, particularly fit indices. For example, a sample of 194 papers published by the American Psychological Association found that over 75% of articles that contain confirmatory factor analyses report the CFI and RMSEA (Jackson et al. 2009). Still, there was significant variability with the reported fit indices, with 34% reporting the Goodness of Fit Index (GFI), 23% reporting the NFI, and 46% reporting the TLI. Thus, the evidence suggests that journals may need more standardization of fit index reporting. In addition, journals have a responsibility to prevent the publication of articles created using unethical research practices. Still, some might argue that it is not the journal's responsibility to assure that their articles follow a standard of reporting fit indices. Instead, some might argue that it is the reviewers' responsibility to assure proper reporting practices. As such, the journals may be responsible for ensuring the reviewers are adhering to standard reporting practices.

Second, to limit the potential of selective reporting of fit indices, researchers should cite their method of reporting. Multiple methods of reporting fit indices exist. Some methods suggest that researchers report the same indices, such as Kline (2016), who recommends reporting the model chi-square statistic, RMSEA, CFI, and the SRMR. Some researchers suggest one should report the TLI, CFI, and RMSEA for one-time analyses and then only report other fit indices when making modifications to the model (Schreiber et al. 2006). Some suggest a hybrid, where researchers must always report the model chi-square, SRMR and then choose a parsimonious index and a relative index (Ockey and Choi, 2015). Finally, some allow researchers to choose one absolute, incremental, and parsimonious fit index (Jackson et al. 2009). Thus, researchers have many methods to choose from when reporting fit indices.

Still, there are limitations to selecting a method when reporting fit indices. First, every method has limitations. For example, the Kline method does not allow for parsimonious fit indices, which reveal a worse fit for more complex methods. Jackson et al. (2009) method still allows researchers to select the best-fit indices of the domains of fit indices to report. The method of Ockey and Choi (2015) limits researchers to specific indices and allows for the ability to select the fit indices that estimate a better model fit. Moreover, researchers will have the potential to selectively pick a method *post-hoc* that makes their models appear to fit better. This selective use of methods may reduce the selective reporting of fit indices; however, it does not stop them. Therefore, the problem with fit index reporting is not the fit indices themselves; rather, the problem comes from the intention and motivation of the researchers to misrepresent their data-model fit. Thus, researchers still need to work on ways of refining the standardization of reporting of fit indices.

USING FIT INDICES TO JUSTIFY POORLY FITTING MODELS

The chi-square exact fit test is sensitive to and suggests poor model fit from minor and typically insignificant model misspecifications (Bentler and Bonett, 1980). With sample sizes between 75 and 200, the chi-square test is typically an appropriate indicator of model fit. However, when the sample size is over 400, most models are rejected. This sensitivity to minor model misspecifications limits the utility of the chi-square exact fit test.

The researcher's ability to detect if a model fits the observed data is limited due to the chi-square exact fit test sensitivity to sample size, so researchers typically rely on other fit measures. Some researchers may use fit indices to justify poorly-fitted models. For example, we can imagine that researchers are testing a model using a dataset of 400 observations. Almost certainly, the chi-square test will suggest that the model does not fit the data. In this example, imagine that the chi-square test is very high, given the degrees of freedom and sample size (e.g., $\chi^2(1) = 10{,}000$, $p < 0.001$). The chi-squared test is much higher than expected, even though the test is sensitive to sample size (i.e., the chi-square test suggests that the model does not fit the data even when accounting for being overpowered). Some fit indices for this model might suggest that the model moderately fits (e.g., a CFI of 0.83). The researcher may then ignore the exact fit test and rely on the CFI to justify the model's fit. Further, a fit index may appear to suggest good data-model fit even when a majority of the pattern coefficients are nonsignificant or weak. This pattern of reporting may mislead the readers to believe that some models fit the data better than they appear given the chi-square exact fit test and pattern coefficients. The problem is not with the fit indices (i.e., the fit indices report the information they were designed to report); rather, the problem is when researchers use the fit indices to argue that a model fits the data when there are major areas of misfit.

FUTURE DIRECTIONS

Researchers should consider the three-step process by Kline (2016) for assessing model fit instead of relying on fit indices. Kline (2016) suggested this method to retain a model as one plausible explanation of the data, even when the exact fit test suggests that the specified model does not fit the data. Step 1 involves fitting the model to the data and reporting the exact fit test. If the model passes the exact fit test, then the researcher will temporarily retain the model as one plausible explanation for the data. If the model fails the exact fit test, then the researcher will tentatively reject the model. Step 2 involves examining standardized and correlational residuals. Standardized residuals are a standardized measure of the error between the observed data and the model-implied data for each piece of unique information in the model-implied covariance matrix. The correlational residuals measure the error between the underlying correlation between items and the model-implied correlations between items. Kline recommends that researchers reject the model if there are numerous correlational residuals (associated with significant standardized residuals) with an absolute value of greater than 0.1 and retain the model if there are no significant correlational residuals. This second step means that researchers may reject models that pass the exact fit test and retain models that fail the exact fit test. Step 3 involves reporting the RMSEA, CFI, and SRMR but not using these fit indices to justify the model fit.

The method of Kline (2016) has several benefits over using the exact fit test or fit indices alone. First, the method of Kline (2016) of examining standardized and correlational residuals allows researchers to assess the fit of individual parts of a model instead of the model as a whole. This benefit allows researchers to assess where the model fits poorly and adjust accordingly (i.e., adding an extra parameter). Second, the method of Kline (2016) allows models that have failed the exact fit test to be redeemed. This benefit removes the emphasis on the exact fit test and allows the researcher to assess if the model failed the exact fit test due to large residuals or just minor model misspecifications. These benefits suggest that using the method of Kline (2016) may be more valid than using fit indices alone.

LIMITATIONS OF FIT INDICES

Although fit indices provide helpful information in assessing data-model fit, there are several notable limitations. First, simulation studies suggest that the implications of cut-off values change when loading and sample size are manipulated (Sharma et al. 2005). This research suggests that proper cut-offs for fit indices (i.e., 0.95 for CFI; Schreiber et al. 2006) changes as a function of the strength of the loadings from the common factors to the indicators, making these fit index cut-offs unreliable. Second, fit indices measure the average fit of the model across parameters and do not allow researchers to assess for the fit of different parameters. This limitation implies that a model with suitable fitting and poor fitting parameters may give a similar fit index as a model with average fitting parameters across the model. Finally, fit indices are only one of many methods that assist researchers in assessing data-model fit. For example, instead of relying exclusively on fit indices researchers can use relative fit across multiple competing models and select the model that demonstrates the best fit. Further, researchers may consider not relying only on empirical methods to determine a model's fit, but also instead using theory and logic to determine which models fit better. For example, a model that is weakly justified theoretically but fits the data well (i.e., solely empirically driven) may not be a model of the hypothesized phenomenon that is as valid as a model that does not fit the data as well but has stronger theoretical support (e.g., Box, 1976; Hox and Bechger, 1999). Further, using pluralistic methods over a single method (i.e., fit indices alone), such as the method of Kline (2016), relative fit comparisons, and theoretical justification in addition to fit indices may guard against the misuse of fit indices (Mayrhofer and Hutmacher, 2020; Zitzmann and Loreth, 2021).

CONCLUSION

Fit indices in structural equation modeling provide helpful information about the data-model fit; however, researchers should use fit indices responsibly and ethically to assure that they do not misrepresent the fit of models. The suggestions in the current paper may limit the misuse of fit indices; however,

researchers may still misuse these suggestions. To maintain the credibility of analyses under the structural equation modeling framework, researchers have a responsibility to uphold the standards of reporting set forth by the experts in the field. The ethical use of fit indices sustains the scientific rigor of the social sciences commanded by empirical investigations. Furthermore, the responsible use of structural equation modeling techniques will allow social scientists to build on the existing framework, which may increase the potential to answer more complex and essential questions.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Bentler, P. M., and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88, 588–606. doi: 10.1037/0033-2909.88.3.588

Box, G. E. P. (1976). Science and statistics. *J. Am. Stat. Assoc.* 71, 791–799. doi: 10.1080/01621459.1976.10480949

Cangur, S., and Ercan, I. (2015). Comparison of model fit indices used in structural equation modeling under multivariate normality. *J. Mod. Appl. Stat. Methods* 4, 152–167. doi: 10.22237/jmasm/1430453580

Hox, J., and Bechger, T. (1999). An introduction to structural equation modeling. *Fam. Sci. Rev.* 11, 354–373.

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

Jackson, D. L., Gillaspy, J. A., and Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: an overview and some recommendations. *Psychol. Methods* 14, 6–23. doi: 10.1037/a0014694

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling*, 4th Edn. New York, NY: Guilford Press.

Mayrhofer, R., and Hutmacher, F. (2020). The principle of inversion: why the quantitative-empirical paradigm cannot serve as a unifying basis for psychology as an academic discipline. *Front. Psychol.* 11:596425. doi: 10.3389/fpsyg.2020.596425

Morgan, G., Hodge, K., Wells, K., and Watkins, M. (2015). Are fit indices biased in favor of bi-factor models in cognitive ability research?: A comparison of fit in correlated factors, higher-order, and bi-factor models via Monte Carlo simulations. *J. Intelligence* 3, 2–20. doi: 10.3390/jintelligence3010002

Ockey, G. J., and Choi, I. (2015). Structural equation modeling reporting practices for language assessment. *Lang. Assess. Q.* 12, 305–319. doi: 10.1080/15434303.2015.1050101

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338. doi: 10.3200/joer.99.6.323-338

Sharma, S., Mukherjee, S., Kumar, A., and Dillon, W. R. (2005). A simulation study to investigate the use of cutoff values for assessing model fit in covariance structure models. *J. Bus. Res.* 58, 935–943. doi: 10.1016/j.jbusres.2003.10.007

Yadama, G. N., and Pandey, S. (1995). Effect of sample size on goodness-fit of-fit indices in structural equation models. *J. Soc. Serv. Res.* 20, 49–70. doi: 10.1300/J079v20n03_03

Zitzmann, S., and Loreth, L. (2021). Regarding an "almost anything goes" attitude toward methods in psychology. *Front. Psychol.* 12:612570. doi: 10.3389/fpsyg.2021.612570

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

METHODS
published: 06 December 2021
doi: 10.3389/fpsyg.2021.755102

Sensitivity Analysis in Nonrandomized Longitudinal Mediation Analysis

*Davood Tofighi**

Department of Psychology, University of New Mexico, Albuquerque, NM, United States

Mediation analysis relies on an untestable assumption of the no omitted confounders, which posits that an omitted variable that confounds the relationships between the antecedent, mediator, and outcome variables cannot exist. One common model in alcohol addiction studies is a nonrandomized latent growth curve mediation model (LGCMM), where the antecedent variable is not randomized, the two covarying mediators are latent intercept and slope modeling longitudinal effect of the repeated measures mediator, and an outcome variable that measures alcohol use. An important gap in the literature is lack of sensitivity analysis techniques to assess the effect of the violation of the no omitted confounder assumption in a nonrandomized LGCMM. We extend a sensitivity analysis technique, termed correlated augmented mediation sensitivity analysis (CAMSA), to a nonrandomized LGCMM. We address several unresolved issues in conducting CAMSA for the nonrandomized LGCMM and present: (a) analytical results showing how confounder correlations model a confounding bias, (b) algorithms to address admissible values for confounder correlations, (c) accessible R code within an SEM framework to conduct our proposed sensitivity analysis, and (d) an empirical example. We conclude that conducting sensitivity analysis to ascertain robustness of the mediation analysis is critical.

Keywords: mediation analysis, sensitivity analysis, no omitted confounder assumption, latent growth analysis, structural equation model (SEM)

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Xiao Liu,
University of Notre Dame, United States
Holly Patricia O'Rourke,
Arizona State University, United States

***Correspondence:**
Davood Tofighi
dtofighi@unm.edu

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

Received: *07 August 2021*
Accepted: *12 November 2021*
Published: *06 December 2021*

Citation:
Tofighi D (2021) Sensitivity Analysis in Nonrandomized Longitudinal Mediation Analysis. Front. Psychol. 12:755102. doi: 10.3389/fpsyg.2021.755102

INTRODUTION

Mediation analysisC has become more common in analyzing complex causal chains in health and psychological studies. One common model in alcohol addiction studies (e.g., Moyers et al., 2009; Hartzler et al., 2011; Maisto et al., 2015) is a nonrandomized latent growth curve mediation model (LGCMM; von Soest and Hagtvet, 2011). The LGCMM, as shown in **Figure 1**, hypothesizes that a nonrandomized antecedent variable (pain) influences both mediators (i.e., mean negative affect and monthly rate of negative affect); these mediators, in turn, cause an outcome variable (alcohol use). Further, the antecedent variable can be also randomized (randomized mediation model). A critical, yet untestable assumption in any mediation model, including the LGCMM, is the no omitted confounder assumption (Judd and Kenny, 1981; Robins and Greenland, 1992; Pearl, 2014; MacKinnon and Pirlott, 2015; Valente et al., 2017). A no omitted confounder assumption states that an omitted variable (confounder) may not exist if it confounds the relationships among the antecedent, mediators, and outcome variable. In a randomized mediation model, this assumption implies that, by properly randomizing the antecedent variable, we can rule out the effect of a confounder on the antecedent variable to the mediators and on the outcome variable relationships but not on the mediators to outcome relationships. In a nonrandomized mediation model, however,

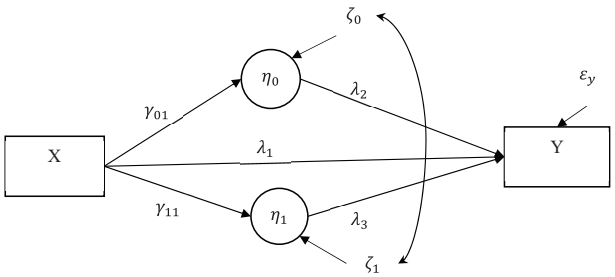

FIGURE 1 | Latent growth curve mediation model (LGCMM). The antecedent variable (*X*) was pain at 4 months after treatment. The two covarying mediators were the latent intercept (η_0 = mean negative mood at 4 months) and slope (η_1 = monthly rate of negative affect) for the repeated measures of negative affect. The outcome variable (*Y*) measures percent drinking days (PDD) at 16 months. *C* denotes a set of covariates (e.g., background variables). A single-headed arrow shows the direct effect of a variable at the origin on the variable at the end of the arrow. A curved double-headed arrow shows covariance between the two residuals.

an omitted confounder can influence *all* the relationships among antecedent variable, mediators, and outcome variable. As a result, it is more challenging to assess the impact of violating the no confounding assumption because the additional patterns of confounding can happen with a nonrandomized mediation model when that model is compared to a randomized mediation model. Because the no omitted confounder assumption is not testable and because the proper randomization of the antecedent and mediator variables is absent, researchers have recommended sensitivity analysis (Imai et al., 2010; VanderWeele and Arah, 2011; Tofighi et al., 2013, 2019; Albert and Wang, 2015; VanderWeele, 2015; Tofighi and Kelley, 2016). A sensitivity analysis assesses the impact of various degrees of violation of the no omitted confounder assumption on the model parameter estimates and on any inferences about the indirect effects.

Despite the prevalence of nonrandomized longitudinal mediation studies in areas such as alcohol addiction (e.g., Moyers et al., 2009; Hartzler et al., 2011; Maisto et al., 2015), most research attention has been on randomized mediation model in both multilevel/longitudinal and single-level data as a means of improving causal inference. In fact, to our knowledge, no study to date has offered a method to conduct sensitivity analysis for a nonrandomized longitudinal mediation model with two covarying mediators in a structural equation model (SEM), a multivariate framework to study covariance structure. Previous research mostly focused on the sensitivity analysis for a randomized model with a single-level data in SEM or in a potential outcome framework (Imai et al., 2010; Cox et al., 2013; Albert and Wang, 2015; Valente et al., 2017; Hong et al., 2018; Lindmark et al., 2018; McCandless and Somers, 2019). For a mediation model with two independent mediators, Imai and Yamamoto (2013) studied sensitivity analysis and strongly assumed independence between the two mediators; their technique, however, cannot be directly applied to a model with covarying mediators as it could result in bias in estimating indirect effects (VanderWeele, 2015). Several studies proposed randomized and nonrandomized sensitivity analysis for sequential mediation models, where one mediator is assumed to sequentially cause another mediator, in both randomized (Imai and Yamamoto, 2013; Daniel et al., 2015) and nonrandomized models (Harring et al., 2017). Because of the strong assumption that the mediators are measured in chronological order, the model specification, interpretation, and sensitivity analysis techniques developed for a sequential mediation model are not directly applicable to a model with covarying mediators, where the mediators freely covary but do not causally impact one another. In multilevel/longitudinal mediation analysis, methods to conduct sensitivity analysis have been proposed for nonrandomized (Tofighi and Kelley, 2016) and for randomized single-mediator model (Bind et al., 2016; Talloen et al., 2016). For two mediators, Tofighi et al. (2019) proposed an SEM-based sensitivity analysis method for a randomized LGCMM. However, Tofighi et al. (2019) did not consider a nonrandomized model where a confounder can influence a pair of relationships among antecedent variable, mediators, and outcome variable. To our knowledge, no study to date has extended sensitivity analysis to a nonrandomized LGCMM in an SEM framework.

Nonrandomized LGCMM poses interwoven challenges compared to either single-level or longitudinal randomized models. First, nonrandomization means a confounder can impact the antecedent as well as the mediators and the outcome variable. Thus, a confounder may impact not only the relationships between the mediators and the outcome variable (as in a randomized mediation model) but also may affect additional relationships of the antecedent to each mediator variable and to the outcome variable. This potential influence poses two additional challenges. The first issue is how to model and estimate biasing impact of a confounder on the antecedent variable in an SEM framework if the antecedent variable is exogenous. In a situation where an antecedent variable is exogenous, the covariates, if they exist, do not influence the antecedent variable. This is a critical issue to address in sensitivity analysis because, in an SEM framework for mediation analysis, an antecedent variable (randomized or not) is generally modeled as an exogenous (and fixed) rather than an endogenous (and random) variable when the covariates are not assumed to influence the antecedent variable. The challenge is to propose a method to convert the antecedent variable without a predicting covariate that is modeled as an exogenous variable into an endogenous variable so that potential impact of omitted confounder on the antecedent variable can be modeled through a confounder correlation.

The second challenge arises because of the additional relationships that can be influenced by a confounder in a nonrandomized model compared to a randomized model. In this instance, conceptualizing, estimating, and interpreting all the confounding relationships and their impacts on the indirect effect estimates will be more complicated than any other sensitivity analysis that has been performed. Also, in a longitudinal model, the repeated measures are correlated; thus, special techniques such as multilevel modeling (Raudenbush and Bryk, 2002) are required to account for lack of independence and to make correct inference about uncertainty of the parameter estimates. In

addition, because of the multilevel structure of data, confounders can impact the model variables at different levels of aggregation (Tofighi and Kelley, 2016), and, thus, techniques developed for a single-level model may not be directly applicable. Further, the existence of two covarying mediators requires that the indirect effect through each mediator be simultaneously estimated. Conducting sensitivity analysis for each mediator separately while ignoring the other covarying mediators, as is done in a single-mediator model, is likely to result in bias because the two mediators are covarying (VanderWeele, 2015). Thus, techniques developed for a single-mediator model cannot be directly used to conduct sensitivity analysis in a two covarying mediator model. Lastly, given a variety of patterns of confounding bias, summarizing the impact of confounding bias succinctly enables researchers to assess sensitivity of the parameter estimate as well as statistical inference to the confounding bias. Given the importance of nonrandomized longitudinal model in practice and the unresolved practical and theoretical issues that hinder conducting sensitivity analysis for such models, proposing a method on how to conduct sensitivity analysis that can be implemented in SEM framework using available software packages is critical.

In this paper, we extend a sensitivity analysis technique from a randomized longitudinal mediation model to a complex, nonrandomized longitudinal mediation model, such as the model illustrated in **Figure 1**, in an SEM framework. More specifically, we extend a technique, termed *correlated augmented model sensitivity analysis* (CAMSA), that was developed for a randomized longitudinal mediation model to conduct sensitivity analysis in nonrandomized longitudinal mediation model with two covarying mediators (Tofighi et al., 2019). The extended CAMSA augments a nonrandomized mediation model with confounder correlations induced by a hypothesized confounder and addresses the unresolved challenges mentioned previously in modeling the biasing impact of the confounder bias. We present analytic results showing the confounder correlations are a function of omitted confounder relations to the model variables; we thereby show how the confounders correlations can be used to estimate confounding biases. We further present results on how to model and estimate confounder correlations in a nonrandomized longitudinal mediation model using the lavaan package (Rosseel, 2012), an opensource, freely available SEM package within the R statistical computing software framework (R Development Core Team, 2020). We present R code along with detailed instruction and an empirical example on how to conduct, interpret, and present the results of this proposed sensitivity analysis[1].

SENSITIVITY ANALYSIS FOR NONRANDOMIZED MEDIATION MODEL

In this section, we extend CAMSA to conduct sensitivity analysis for LGCMM (**Figure 1**). For simplicity but without loss of generality, we consider an LGCMM without any covariates. However, the results presented in this section will hold when adding the covariates to LGCMM in **Figure 1** as we will demonstrate in the empirical example section. A crucial step in CAMSA is to specify a set of confounder correlations; a confounder correlation is used to model the impact of the omitted confounders on the model parameters. More specifically, confounder correlations are specified between the residuals associated with the endogenous variables (i.e., the variables with arrows pointed toward them). In extending CAMSA to the nonrandomized LGCMM, we faced several challenges. First, we needed to determine how to specify confounder correlations between an antecedent variable, which is an exogenous variable with no residual term, and the residuals associated with the endogenous variables in the model. This step is required because CAMSA uses confounder correlations to model biasing confounder effects. Second, we found a lack of clarity about whether the confounder correlations are uniquely a function of the confounder effects on the model parameters or whether they are also a function of the existing relationships between the variables. Clarifying such relationships would elucidate what confounder correlations are modeling. Thus, it is necessary to analytically demonstrate relationships between the confounder correlations and the effect of the confounder on the model parameters. Third, because of an infinite number of the combinations of confounder correlations, we were challenged to find plausible values for confounder correlations that are admissible and practical. In finding admissible values, we employ and implement different methods such that the correlation matrix is positive definite. To find sensible values, we propose steps to explore a finite set of plausible confounder correlation patterns as opposed to examining an infinite number of patterns; thus, we provide a more accessible way for researchers to conduct and interpret sensitivity analysis when facing an infinite number of choices for confounder correlations. In the next section, we introduce and extend CAMSA to the nonrandomized LGCMM. We then present formulae for computing confounder correlations in CAMSA. Next, we show equivalency between the correlated augmented model and a latent augmented model, an LGCMM that models a confounder explicitly. Finally, we present methods to generate admissible confounder correlations for CAMSA.

CAMSA

To conduct CAMSA and address the challenges outlined above, we first specify the correlated augmented model, which adds to the original LGCMM (**Figure 1**) with the confounder correlations as shown in **Figure 2**. But first, we address the challenge that, in the model in **Figure 1**, the antecedent variable is not endogenous variable but is an exogenous variable without a residual term. To solve this issue, we introduce a residual term ε_{xi} for the antecedent variable X. To ensure that the model is identified, we fix variance of the residual term to equal the variance of the antecedent variable. The residual term would explicitly specify the antecedent variable as an endogenous variable as opposed a "fixed" exogenous variable that tends to be a default setting in SEM software packages. That is, specifying this residual term

[1]The R script for our proposed CAMSA for the empirical example is provided in the online **Supplementary Material**.

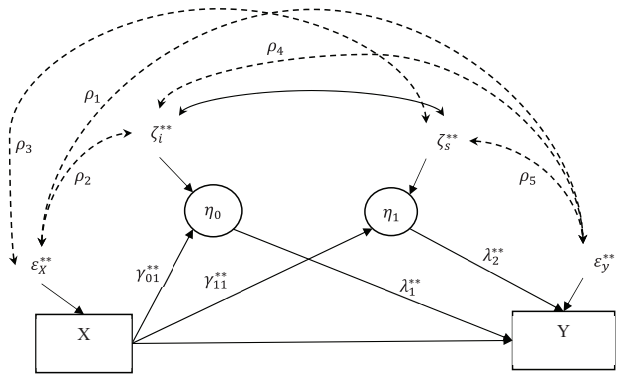

FIGURE 2 | Correlated augmented LGCMM. The antecedent variable (*X*) was pain at 4 months after treatment. The two covarying mediators were the latent intercept (η_0 = mean negative mood at 4 months) and slope (η_1 = monthly rate of negative affect) for the repeated measures of negative affect. The outcome variable (*Y*) measures percent drinking days (PDD) at 16 months. A single-headed arrow shows the direct effect of a variable at the origin on the variable at the end of the arrow. A solid double-headed arrow shows covariance between the two residuals. A dashed doubled headed arrow shows confounder covariance (correlation) between the residuals.

will convert the role of the antecedent variable from exogenous to endogenous, thus permitting us to specify the confounder correlations between the antecedent variable residual term and the other residual terms in the model. Now, we can model confounder bias as the confounder correlations among all the residual terms of the endogenous variables. In addition, by converting the antecedent variable to an endogenous variable, we can manipulate elements of the covariance matrix of endogenous variables. Below we present the equations for this model where Eq. (1) demonstrates specification for converting the antecedent variable into an endogenous variable. The superscript "**" denotes the parameters for the correlated augmented model[2].

$$x_i = \alpha_3^{**} + \varepsilon_{xi}^{**} \tag{1}$$

$$m_{ij} = \eta_{i0} + \eta_{i1}\, t_{ij} + \; e_{ij}^{**} \tag{2}$$

$$\eta_{i0} = \alpha_0^{**} + \gamma_{01}^{**}\, x_i + \; \zeta_{0i}^{**} \tag{3}$$

$$\eta_{i1} = \alpha_1^{**} + \gamma_{11}^{**}\, x_i + \; \zeta_{1i}^{**} \tag{4}$$

$$y_i = \alpha_2^{**} + \lambda_1^{**} x_i + \; \lambda_2^{**}\eta_{i0} + \lambda_3^{**}\eta_{i1} + \varepsilon_{yi}^{**} \tag{5}$$

In the above equations, subscript *i* denotes person $i = 1, \ldots, N$, and subscript *j* denotes an occasion for the repeated measures variable m_{ij} where $j = 1, \ldots, p$. Variables m_{ij} and t_{ij} are the repeated measures on the observed mediators and the time score, respectively; y_i is the outcome variable, and x_i is the antecedent variable. Latent intercept and slope are denoted by η_{i0} and η_{i1}. The terms α_0^{**} and α_1^{**} denote the intercepts for the latent growth factors. The intercept for the outcome variable is α_2^{**}; the intercept for the antecedent variable is α_3^{**} where it is, in fact, estimated by the sample mean for the antecedent variable. The parameters γ_{01}^{**} and γ_{11}^{**} quantify the effects of the antecedent variable on the latent intercept and slope, respectively. The regression coefficients $\lambda_1^{**}, \lambda_2^{**}$, and λ_3^{**} quantify the effects of the antecedent variable, latent intercept, and slope on the outcome variable, respectively.

The second part of positing the correlated augmented model is to specify the variances and covariances between the residuals for the model. From a multilevel perspective, two levels of residuals for LGCMM exist. First, there are *p* Within (Level-2) residuals associated with repeated measures m_{ij}s, $\varepsilon_W^{**} = \left(e_{i1}^{**}, \ldots, e_{ip}^{**}\right)^T$, where T denotes vector transpose operator. Second, there are four Between (Level-2) residuals $\varepsilon_B^{**} = \left(\varepsilon_{ix}^{**},\ \varepsilon_{iy}^{**},\ \zeta_{0i}^{**},\ \zeta_{1i}^{**}\right)^T$, where ε_{ix}^{**}, ε_{iy}^{**}, ζ_{0i}^{**}, and ζ_{1i}^{**} are associated with the antecedent variable, outcome variable, and latent intercept and slope, respectively. Note that ε_{ix}^{**} is included in vector of the residuals because we explicitly specify the antecedent variable as an endogenous variable with a residual term. We assume that the covariances between the Level-1 and Level-2 residuals to be zero (Raudenbush and Bryk, 2002).

For each level, the vector of residuals has the multivariate normal distribution with a mean vector of zero and a covariance matrix. For the Within residuals, the upper-triangle covariance matrix is:

$$\Sigma_W^{**} = \begin{bmatrix} \sigma^2_{e_1^{**}} & 0 & 0 \\ & \ddots & 0 \\ & & \sigma^2_{e_p^{**}} \end{bmatrix}$$

where $\sigma^2_{e_1^{**}}$ and $\sigma^2_{e_p^{**}}$ are the residual variance for m_{i1} and m_{ip}, respectively. We assume that the covariances between the repeated measures are zero although one could estimate the Within residuals when this premise is supported by theory. For the Between residuals, the upper-triangle covariance matrix is:

$$\Sigma_B^{**} = \begin{bmatrix} \sigma^2_{\varepsilon_{xi}^{**}} & \sigma_{\varepsilon_{xi}^{**}}\sigma_{\varepsilon_{yi}^{**}}\rho_1 & \sigma_{\varepsilon_{xi}^{**}}\sigma_{\zeta_0^{**}}\rho_2 & \sigma_{\varepsilon_{xi}^{**}}\sigma_{\zeta_1^{**}}\rho_3 \\ & \sigma^2_{\varepsilon_{yi}^{**}} & \sigma_{\varepsilon_{yi}^{**}}\sigma_{\zeta_0^{**}}\rho_4 & \sigma_{\varepsilon_{yi}^{**}}\sigma_{\zeta_1^{**}}\rho_5 \\ & & \sigma^2_{\zeta_0^{**}} & \sigma_{\zeta_0^{**},\zeta_1^{**}}\rho_{\eta_0,\eta_1} \\ & & & \sigma^2_{\zeta_1^{**}} \end{bmatrix}.$$

As shown in **Figure 2**, ρ_1, ρ_2, and ρ_3 are the confounder correlations between the antecedent variable and outcome variable, the latent intercept and slope, respectively; ρ_4 and ρ_5 are the confounder correlations between the latent intercept and slope and the latent slope and outcome variable, respectively. The terms σ^2s denote the variances of the respective residuals.

The confounder correlations ρ_1 to ρ_5 are assumed to model the effect of the omitted confounder bias on the model parameters although the exact nature of the relationships between the confounder correlations and the omitted confounder remains unclear. Note that if we had assumed that all the confounders were included in the model (e.g., the covariates included the confounders), most if not all the confounder correlations would equal zero (Tofighi et al., 2013, 2019; Tofighi and Kelley, 2016).

[2]Note that for simplicity and without loss of generality, we consider an LGCMM without any covariates. However, the results presented in this section hold when adding the covariates to LGCMM **Figure 1** as will be shown in the "Empirical Example" section.

If all the confounders were included in the model, except for the residuals of the latent intercept and slope, the residuals associated with the antecedent variable, latent intercept and slopes, and the outcome variable would not correlate merely because of the omitted confounders[3]. This argument will be made clearer when we show the relationships between the confounder correlations and the omitted confounder effects later in this article. The covariance of the residuals between the latent intercept and slope is usually freely estimated (Singer and Willett, 2003). In general, the covariance between the intercept and slope should not be fixed at zero because of its potential substantive interpretation (Snijders and Bosker, 2011). As we will discuss later, the covariance between the intercept and slope could be biased because of the confounder bias when ρs are non-zero.

Before conducting CAMSA using the correlated augmented model, three important issues must be addressed. First, we need to derive analytical formulas to transform the confounder correlations into confounder covariances that are covariances between the residuals quantifying the effects of the omitted confounders. Second, we need to determine the relationships between the confounder correlations and the effect of the omitted confounder on the model parameters. Third, we need methods to generate admissible confounder correlation values that are of substantive interest. We address these issues in the next sections.

Transforming Confounder Correlations Into Confounder Covariances

In conducting CAMSA, we cannot use confounder correlations directly to specify a correlated augmented mediation model in SEM framework because of scaling of the endogenous variables. Rather, we need to use the confounder covariance and estimates of the residual variance and then convert the fixed values of confounder correlations to confounder covariance using the derived computational formulas. We use the derived formulas to estimate the correlated augmented model (see **Supplementary Material** for details on deriving the formulas).

$$cov(x_i, \sigma_{\varepsilon_{yi}^{**}}) = \rho_1 \sigma_{\varepsilon_{xi}^{**}} \sigma_{\varepsilon_{yi}^{**}}$$

$$cov(x_i, \zeta_{0i}^{**}) = \rho_2 \sigma_{\varepsilon_{xi}^{**}} \sigma_{\zeta_{0i}^{**}}$$

$$cov(x_i, \zeta_{1i}^{**}) = \rho_3 \sigma_{\varepsilon_{xi}^{**}} \sigma_{\zeta_{1i}^{**}}$$

$$cov(\zeta_{0i}^{**}, \varepsilon_{yi}^{**}) = \rho_4 \sigma_{\zeta_{0i}^{**}} \sigma_{\varepsilon_{yi}^{**}}$$

$$cov(\zeta_{1i}^{**}, \varepsilon_{yi}^{**}) = \rho_5 \sigma_{\zeta_{1i}^{**}} \sigma_{\varepsilon_{yi}^{**}}$$

Equivalence Between Correlated Augmented Model and Latent Augmented Model

In this section, we show the equivalence between the correlated augmented model used in CAMSA and the *latent augmented model* in **Figure 3**. We use the term *latent augmented model* because ϖ, termed a *latent confounder*, denotes a latent variable that accounts for a linear combination of the potential omitted confounders. Establishing equivalency is critical because it is unclear whether the confounder covariances/correlations in the correlated augmented model exclusively account for the confounder correlations or for correlations not caused by an omitted confounder. Model specification and equations for the latent augmented model and detailed analytic proofs of the equivalency between the latent augmented model and correlated augmented model are shown in **Supplementary Material**.

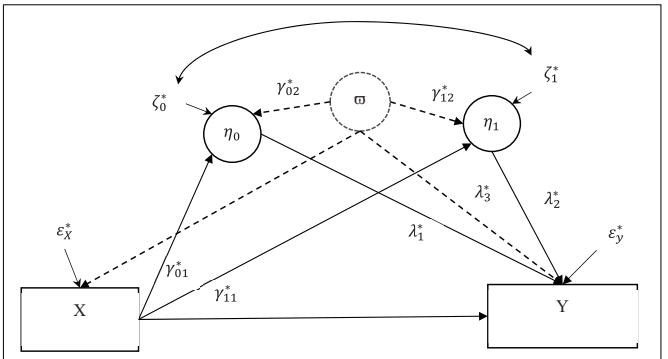

FIGURE 3 | A latent augmented LGCMM. The antecedent variable (*X*) was pain at 4 months after treatment. The two covarying mediators were the latent intercept (η_0 = mean negative mood at 4 months) and slope (η_1 = monthly rate of negative affect) for the repeated measures of negative affect. The outcome variable (*Y*) measures percent drinking days (PDD) at 16 months. A solid single-headed arrow shows the direct effect of a variable at the origin on the variable at the end of the arrow. A solid double-headed arrow shows covariance between the two residuals. The dashed circle shows the latent proxy variable ϖ and the dashed arrows show the confounder parameters modeling the effect of the latent proxy variable on the endogenous variables.

A significant contribution of our paper is to establish that the latent augmented model is equivalent to the correlated augmented model. That is, there is a one-to-one relationship between the corresponding parameters from the two model. To confirm this correspondence, we must show that the confounder correlations/covariances in the correlated augmented model are, in fact, functions of the confounding parameters in the latent augmented model. It is not trivial that the confounder correlations/covariances specifically model the effects of confounders and not the other relationships between the variables in the model. For example, we show that the covariance between the latent intercept and slope in the correlated augmented model is not, in general, equal to the corresponding covariance between the intercept and the slope in the latent augmented model.

Using a latent augmented model in sensitivity analysis is an extension of the sensitivity analysis technique used in randomized LGCMM (Tofighi et al., 2019) and multilevel SEM (Tofighi and Kelley, 2016) and is similar to the phantom variable technique used in single-level SEM (Harring et al., 2017). We only use the latent augmented model to exhibit

[3]The residuals could be correlated due to other factors, such as common methods factors (Podsakoff et al., 2003).

equivalency but not to conduct sensitivity analysis because using the latent augmented model over the correlated augmented model potentially produces negative residual variance (Tofighi et al., 2019). Furthermore, using a latent augmented model means that the confounder parameters, regression coefficients measuring confounding bias on the model variables, are not easily interpretable because of scaling of the variables. These issues are remedied in the correlated augmented model because we use confounder correlations, which are effect size measures, and thus are more easily interpreted to gauge the impact of confounding bias.

Generating Confounder Correlation Matrix

In the correlated augmented model, the confounder correlations are a set of fixed values that researchers establish with some restrictions that will make correlation values admissible. We will discuss different methods of finding admissible confounder correlation values. One important contribution of our model is that we propose a two-step procedure to investigate values for the confounder correlations that has not been used in sensitivity analysis literature. The two-step procedure uses the following two methods of generating admissible correlation values: (a) Toeplitz matrix method and (b) nearest positive-definite (PD) matrix method. We explain in detail how each method works and enumerate the pros and cons of each method. We show application of the two-step procedure using an empirical example in the next section.

A critical issue that needs to be addressed before conducting CAMSA is finding admissible values for the confounder correlations. Finding correlation confounder values is challenging because the correlations have a restricted range of $[-1, 1]$ and are restricted by the values of other confounder correlations. In other words, the correlation values are not independent. This dependency means that we cannot pick values for a correlation independent of the values of other correlations. For example, for a triplet of correlations, all the values must satisfy the following constraint (Rousseeuw and Molenberghs, 1994):

$$\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2 - 2\rho_{12}\rho_{13}\rho_{23} \leq 1$$

Finding the values of the triplets of correlations that would satisfy the above restriction is not straightforward, especially as the number of correlations increase. This challenge can be more readily seen when arranging correlation values into a matrix as the Between confounder correlation matrix shown below.

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ & 1 & \rho_4 & \rho_5 \\ & & 1 & \rho_{\eta_0,\eta_1} \\ & & & 1 \end{bmatrix} \tag{6}$$

The correlation matrix in (6) must be positive semi-definite (PSD). A square symmetric matrix is PSD if and only if the matrix determinant is greater than or equal to zero. Finding the determinant and setting it to be greater than (equal to) zero would provide us with necessary and sufficient condition for the correlation matrix. However, generating a matrix that would satisfy these conditions is not straightforward. Additionally challenging is generating a symmetric PSD whose diagonal elements are one and off-diagonal elements are between -1 and 1 where the correlation values substantively meaningful. The challenges are to find values and patterns of confounder correlations that are of substantive interest while satisfying the PSD condition. Below we discuss a two-step procedure for generating PSD correlation matrices.

Step 1: Toeplitz Matrix Method

One way to generate a correlation matrix is to use a special type of symmetric Toeplitz matrix (Schott, 1997; Wicklin, 2015) in which the main diagonal and diagonals parallel to the main diagonal are constant. We focus on the symmetric Toeplitz matrix whose diagonal elements are one. Consider n real numbers $a_0, a_1, ..., a_{n-1}$ where $a_0 = 1$. Then we can denote a symmetric Toeplitz matrix whose first row is $a_0, a_1, ..., a_{n-1}$ by

$$T_n = T_n[a_0, a_1, \ldots, a_{n-1}] = \begin{bmatrix} a_0 & a_1 & \cdots & a_{n-1} \\ & a_0 & \cdots & a_{n-2} \\ & & \ddots & \vdots \\ & & & a_0 \end{bmatrix} \tag{7}$$

We use the results by Bogoya et al. (2012) to generate a special case for Toeplitz matrix where the matrix is square and symmetric with the main diagonal of one. A main result of Bogoya et al. (2012) is that a symmetric Toeplitz matrix whose first row is a linearly decreasing sequence (i.e., a sequence that decreases by the same amount each time) of non-negative values is PD. Thus, the process creates a sequence of decreasing positive values that generates a correlation matrix (Wicklin, 2015). To illustrate, consider a general polynomial sequence of the form $c_1 - (j-1)c_2$ where c_1 and c_2 are constant and j is the index for column number, $j = 1, ..., n$, — although one could also build a Toeplitz matrix with the first column and then use a row index. The sequence is expanded as follows: $c_1, c_1 - c_2, c_1 - 2c_2, ...$ A necessary condition for the matrix to be PD is that the sequence should positive, thus $c_1 - (j-1)c_2 > 0$. For a correlation matrix, we set $c_1 = 1$ because the main diagonal elements of a correlation matrix equal one. Thus, the condition $1 - (j-1)c_2 > 0$ means that decrement c_2 must satisfy $c_2 < 1/(n-1)$ and that the largest value for j is the dimension of the matrix, n. For example, a decrement for a triplet correlation Toeplitz matrix must be $c_2 < 1/2$. For the triplet of correlation values, when $c_2 = 1/4$, the first row is $(1, 3/4, 1/2)$, and the resulting correlation matrix is

$$\begin{bmatrix} 1 & 3/4 & 1/2 \\ & 1 & 3/4 \\ & & 1 \end{bmatrix}.$$

To generate this matrix in R, we use

```
(T1 <- toeplitz(c(1,3/4, 1/2))) # generates
 a Toeplitz matrix given the first row

##      [,1] [,2] [,3]
## [1,] 1.00 0.75 0.50
## [2,] 0.75 1.00 0.75
## [3,] 0.50 0.75 1.00

det(T1) # determinant of T1

## [1] 0.1875
```

What about the Toeplitz matrix for 4×4 confounder correlation matrix for our LGCMM? The first row for the Toeplitz for the confounder correlation matrix is $(1, 1 - c_2, 1 - 2 \times c_2, 1 - 3 \times c_2)$ where $c_2 < 1/3$. We wrote an R function to generate the first row and used Toeplitz function in R to generate the Toeplitz confounder correlation matrix.

```
first_row <- function(x) {
  if (x >= 1 / 3)
    stop("Choose a value smaller than 1/4!")
  return(c(1, 1 - x, 1 - 2 * x, 1 - 3 * x))
}
(T2 <- toeplitz(first_row(1/4.1)))

##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.7560976 0.5121951 0.2682927
## [2,] 0.7560976 1.0000000 0.7560976 0.5121951
## [3,] 0.5121951 0.7560976 1.0000000 0.7560976
## [4,] 0.2682927 0.5121951 0.7560976 1.0000000

det(T2)

## [1] 0.07360849
```

So far, we have talked about generating the confounder correlation matrix using Toeplitz matrix and a result from Bogoya et al. (2012) that only generates positive confounder correlations. What if we would like to have negative confounder correlations as well? Bogoya et al. (2012) extended their results to include the negative correlation by showing that a linearly decreasing sequence of numbers can include negative values if the sum of the values in the sequence remains positive, that is,

$$\sum_{j=1}^{n} c_1 - (j-1)c_2 = nc_1 - c_2 \sum_{j=1}^{n} (j-1)$$

$$= nc_1 - c_2 \frac{n(n-1)}{2} > 0.$$

Given the $c_1 = 1$ for a Toeplitz correlation matrix, we have

$$n - c_2 \frac{n(n-1)}{2} > 0$$

$$2 - c_2(n-1) > 0$$

$$c_2 < \frac{2}{(n-1)}$$

The above result indicates that, if we want to create a Toeplitz correlation matrix with both positive and negative confounder correlations, then we must choose a decrement that satisfies the condition $c_2 < \frac{2}{(n-1)}$. However, if we want to generate a correlation matrix with positive values, the decrement must satisfy this condition $c_2 < \frac{1}{(n-1)}$. Note that the decrement for the positive confounder correlation is smaller than one for positive and negative confounder correlation. Now, we modify our R function to indicate that one may choose a decrement that would generate a Toeplitz matrix that includes both positive and negative correlation matrix:

```
first_row <- function(x) {
  if (x >= 2 / 3)
    stop(
      "Choose a value smaller than 1/2 for both
positive and negative correlation. Choose a a
value smaller than 1/4 for only positive
correlation."
    )
  return(c(1, 1 - x, 1 - 2 * x, 1 - 3 * x))
}
(T3 <- toeplitz(first_row(1/3.9)))

##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.7435897 0.4871795 0.2307692
## [2,] 0.7435897 1.0000000 0.7435897 0.4871795
## [3,] 0.4871795 0.7435897 1.0000000 0.7435897
## [4,] 0.2307692 0.4871795 0.7435897 1.0000000

det(T3)

## [1] 0.08299326
```

Using this algorithm provides a great flexibility in choosing a select number of confounder correlations to examine a relatively wide range of the indirect effect values as well as to examine model convergence. Thus, we recommend this algorithm be used as an initial step to inspect the correlation confounders and indirect effects values as well as the non-convergence of the mediation model. While relative simplicity of this method and the relationship between the confounder correlations dictated by Toeplitz algorithm are advantages, the algorithm also limits the range of confounder values because the confounder correlation values follow the Toeplitz algorithm. Thus, we recommend researchers use this initial step to investigate confounder correlation values and their impacts on the sensitivity of the indirect effects as well as on the model convergence.

Step 2: Nearest Positive-Definite Method

In the second step, we examine in more depth the range of confounder correlation values that led to convergence of the model using the relevant information from Step 1. Examining the range of possible values would exhaust the memory and computational resources. Even a more limited range of confounder correlations could take hours on faster available PCs. Given that we will generate thousands of correlation matrices, using the confounder correlation values from the initial phase will help us focus on the select ranges of the confounder

correlations and, thus, be able to examine more thoroughly the combination of confounder correlation values within the select ranges gleaned from Step 1.

In this second step, we use the range of values that do not lead to non-convergence of the model to generate many correlation matrices to be used in sensitivity analysis. However, as mentioned before, not all combinations of the correlation values would lead to PD confounder correlation matrices. To solve this problem, we use an algorithm suggested by Higham (2002) to transform a non-PD correlation matrix into a "nearest" PD matrix. The nearest PD matrix is achieved by repeatedly projecting the original non-PD matrix onto the set of all symmetric positive semidefinite matrices (termed *cone*) with unit diagonal entries.

The benefit of Higham's (2002) algorithm is that we can choose and control the range of values for each confounder correlation. For example, we can choose a few values for ρ_{XY}, such as small, medium, and large according to Cohen's (1988) guideline, while we choose a continuous range for other confounder correlations, for example, $0 \leq \rho_{XM1} \leq 0.5$. A limitation of this method is that the many combinations of the confounder correlation values can lead to non-convergence of the model. As a result, this method can be computationally expensive even with the computational power of the modern computers. Further, this method is a compromise between having more control over the range of the confounder correlation values and the convergence rate of the model. Next, we illustrate an application of our proposed CAMSA to an empirical example and show that CAMSA is generalizable to a nonrandomized LGCMM with covariates.

EMPIRICAL EXAMPLE

To illustrate the application of CAMSA to a nonrandomized longitudinal growth model, we used data from Combined Pharmacotherapies and Behavioral Interventions for Alcohol Dependence study (COMBINE; The COMBINE Study Research Group, 2003), a randomized control trial that studied 16 weeks of active treatment alcohol use disorder on 1,383 participants recruited across 11 sites. The participants received nine individual treatments or a combination of the following treatments: sobriety and enhance medication adherence training (Medical Management, MM), individualized psychotherapy for outpatient alcohol dependence (combined behavioral intervention, CBI), medications to reduce alcohol dependency (e.g., acamprosate, naltrexone, or combination of the two), or a placebo. Background information (covariates) prior to treatment and assessment measures at the beginning (baseline).

In our example, we were interested in whether the negative effect of pain on a participant's drinking outcome would be mediated through negative affect. The antecedent variable was pain at 4 months (the end of the treatment). Pain was measured by two items. One item, selected from the 26-item World Health Quality of Life assessment (World Health Organization, 1997), asks "To what extent do you feel that physical pain prevents you from doing what you need to do?" The possible responses range from 1 "not at all" to 5 an "extreme amount." The second question, selected from the 12-item Short Form Health Survey (Ware et al., 1996), asks "During the past 4 weeks, how much did pain interfere with your normal work including both work outside the home and housework?" Again, the possible responses range 1 "not at all" to 5 "extremely." The outcome variable was percent drinking days (PDD) at 16 months and was measured via Form 90 (Miller, 1996).

The two mediators were the intercept and slope for repeated measures of negative affect. Negative affect was measured by the self-reported, 53-item Brief Symptom Inventory (BSI) that measures distress (Derogatis and Melisaratos, 1983). An example item asks, "How much were you distressed by nervousness or shakiness inside?" with responses ranging from 0 "not at all" to 4 "extremely." The BSI was measured at 4 months (the end of treatment), 6.5, 13, and 17 months. For the LGCMM, the latent intercept measured the mean negative affect at 4 months while the latent slope measured the monthly rate of negative affect.

For this example, we also controlled for the following covariates measured at or prior to the baseline: demographics (i.e., gender, marital status, employment status, income, and minority status), baseline alcohol dependence severity (Skinner and Allen, 1982), number of alcohol dependence symptoms (American Psychiatric Association, 1995), readiness to change (DiClemente and Hughes, 1990), and alcohol abstinence self-efficacy (DiClemente et al., 1994). Our proposed CAMSA results are generalizable to LGCMM with the covariates. That is, the analytical results for confounder correlation conversion formulas and equivalency still hold. One notable adaptation when adding covariates is that, because we control for the covariates for antecedent variable as well as the mediators and outcome variable, the antecedent variable is automatically endogenous. Thus, we did not need to explicitly specify the antecedent variable as an endogenous variable.

We fitted LGCMM using lavaan (Rosseel, 2012) and conducted CAMSA in R (R Development Core Team, 2020), an open source, freely available statistical software[4]. If the no-omitted-confounder assumptions hold, the indirect effect through the intercept was 0.096 ($SE = 0.014$), 95% CI [0.07, 0.125] and the indirect effect through the slope was −0.013 ($SE = 0.012$), 95% CI [−0.039, 0.011]. Recall that the latent intercept was the mean negative mood at 4 months and latent slope was the monthly rate of negative affect. These results indicate that pain increased the mean of negative mood at 4 months and that, in turn, increased PDD at 16 months. However, pain does not appear indirectly to change PDD through the negative mood monthly change.

Given that the no-omitted confounder assumption is not testable, we used our proposed method to conduct sensitivity analysis. An important step was to find correlation confounder values that were both feasible and practical. To do that, we followed the steps of our proposed method. In Step 1, we generated structured Toeplitz correlation matrices using the algorithm by Bogoya et al. (2012). We then augmented the model with confounder correlation values, ran the model, and

[4]The R script for our proposed CAMSA for the empirical example is provided in the online **Supplementary Material**.

computed the indirect effect estimates for each model. To save space, we only show a few select combinations of confounder correlations and the indirect effect through latent intercept and slope in **Tables 1**, **2**, respectively; more complete tables containing the results can be found in the **Supplementary Materials**. We found that not all combinations of confounder correlation values would result in convergence. Confounder correlation values equal or greater than the values $\rho_{XY} \geq -0.032$, $\rho_{XM_1} \geq 0.656$, $\rho_{XM_2} \geq 0.312$, $\rho_{M_1M_2} \geq 0.656$, $\rho_{M_1Y} \geq 0.312$, $\rho_{M_2Y} \geq 0.656$ resulted in nonconvergence. Nonconvergence means that the confounder correlation values caused model nonconvergence, and, thus, these confounder correlations were inadmissible. Although these values were proper confounder correlation values from the standpoint of the confounder correlation matrix being PD, the values were not compatible with the data and the model implied correlation structure; thus, these values were discarded. If we fix a correlation, or any parameter for that matter, that is not supported by the data and the model, then the chance of model nonconvergence increases (Anderson and Gerbing, 1988).

In Step 2, using the results from Step 1 as a guide, we explored a wider range of confounder correlation values. We used the near PD method, which allowed us to examine more combinations of confounder correlations for the sensitivity analysis. In addition, given that the range of confounder correlation values were not guaranteed to be PD, we used the near PD algorithm to convert a non-PD confounder correlation matrix into its nearest PD matrix. Like Step 1, we then augmented the model with confounder correlation values, ran the model, and computed the two indirect effect estimates for each model. The five confounder correlations took on five values, −0.3, −0.1, 0, 0.1, and 0.3, and resulted in 15,625 combinations. Of 15,625 estimates for each indirect effect, 15,000 (96%) resulted in nonconvergence.

Because of a relatively large number of estimates, we recommend researchers be deliberate in examining indirect effects for the corresponding range of confounder correlation values. We started by examining the results for zero to small effect ($0 < \rho < 0.1$) for the confounder correlations. For the indirect effect through the intercept, examining the range of values for small to zero showed support for the indirect effect to be robust in that indirect effect remained positive; further, the CI limits were all positive except for a few cases shown in **Table 1**. Maximum indirect effects when the confounder correlations were in the zero to small effect range are also shown in **Table 1**. For the medium to large confounder correlations, however, none of the models converged. Nonconvergence results should be interpreted in the context of the select values for the confounder correlations; we could not conclude that all the medium to large confounder correlations would result in nonconvergence.

For the indirect effect through slope when the confounder correlations were zero to small effect range, the sign of the magnitude and inference about the indirect effect CI limits remained unchanged. **Table 2** shows five combinations of the confounder correlations that resulted in the smallest and the largest indirect effects. The indirect effect estimates remained negative, ranging from −0.0167 to −0.0086. The lower limit of the CIs ranged from −0.0487 to −0.0255 while the upper limit ranged from 0.0153 to 0.0082. Because the indirect effect CI contained zero, the indirect effect did not appear to be different from zero when the confounder correlations ranged from zero to small. Finally, we conducted sensitivity analysis when the confounder correlations ranged from medium to large with the values ranging from 0.3 to 0.5 and from −0.5 to −0.3. All the combinations resulted in nonconvergence.

One important feature of our proposed CAMSA is that we can ascertain sensitivity of the model fit to the confounder correlations by examining convergence of the model fit to the data. The nonconvergence results indicate that the correlated augmented model, which consists of the constraints imposed by the confounder correlations along with the implied covariance matrix and mean structure posited by the model, was not supported by the sample data (Anderson and Gerbing, 1988). The estimation algorithm was not able to find the sample estimates that would maximize the likelihood of data given the correlated augmented model. We concluded that the fit of the

TABLE 1 | A sample of sensitivity analysis results for indirect effect through intercept for zero to small confounder correlation.

ρ_{XY}	ρ_{XM_1}	ρ_{XM_2}	$\rho_{M_1M_2}$	ρ_{M_1Y}	ρ_{M_2Y}	Indirect effect	LL	UL
Non-significant indirect effects								
−0.1	0.1	0	−0.1	0.1	0	0.012	−0.00005	0.02447
−0.1	0.1	0	−0.05	0.1	0	0.012	−0.00005	0.02447
−0.1	0.1	0	0	0.1	0	0.012	−0.00005	0.02447
−0.1	0.1	0	0.05	0.1	0	0.012	−0.00005	0.02447
−0.1	0.1	0	0.1	0.1	0	0.012	−0.00005	0.02447
Largest indirect effects								
−0.1	−0.1	0	−0.1	−0.1	0	0.249	0.203	0.295
−0.1	−0.1	0	−0.05	−0.1	0	0.249	0.203	0.295
−0.1	−0.1	0	0	−0.1	0	0.249	0.203	0.295
−0.1	−0.1	0	0.05	−0.1	0	0.249	0.203	0.295
−0.1	−0.1	0	0.1	−0.1	0	0.249	0.203	0.295

These results are from Step 1, where the structured Toeplitz correlations, ρs, were generated using the algorithm by Bogoya et al. (2012). LL, lower limit; UL, upper limit.

TABLE 2 | A sample of sensitivity results for largest and smallest indirect effect through slope for zero to small confounder correlations.

ρ_{XY}	ρ_{XM_1}	ρ_{XM_2}	$\rho_{M_1M_2}$	ρ_{M_1Y}	ρ_{M_2Y}	Indirect effect	LL	UL
Smallest indirect effect								
0.1	0.1	0	−0.1	−0.1	0	−0.0167	−0.0487	0.0153
0.1	0.1	0	−0.05	−0.1	0	−0.0167	−0.0487	0.0153
0.1	0.1	0	0	−0.1	0	−0.0167	−0.0487	0.0153
0.1	0.1	0	0.05	−0.1	0	−0.0167	−0.0487	0.0153
0.1	0.1	0	0.1	−0.1	0	−0.0167	−0.0487	0.0153
Largest indirect effect								
−0.1	0.1	0	−0.1	0.1	0	−0.0086	−0.0255	0.0082
−0.1	0.1	0	−0.05	0.1	0	−0.0086	−0.0255	0.0082
−0.1	0.1	0	0	0.1	0	−0.0086	−0.0255	0.0082
−0.1	0.1	0	0.05	0.1	0	−0.0086	−0.0255	0.0082
−0.1	0.1	0	0.1	0.1	0	−0.0086	−0.0255	0.0082

These results are from Step 1, where the structured Toeplitz correlations, ρs, were generated using the algorithm by Bogoya et al. (2012). LL, lower limit; UL, upper limit.

posited model itself was sensitive to select medium to large values of confounder correlations because the fit of the model was not supported by the sample data. One interpretation of model convergence sensitivity is that the effect of the confounder correlations would severely degrade the fit of the posited model to the sample data to a degree that the model would not be able to be estimated from the sample data. The fit of the posited model appeared to be sensitive to the confounders with medium to large influence on the model. Thus, we could argue that the posited model as a whole (global fit) and the indirect effects (local fit) as a part of the posited model do not appear to be robust to the confounder correlations ranging from medium to large effects.

In summary, it appears that when the confounder correlations were in the zero to small range, the overall model convergence and the two indirect effects through intercept and slope were less sensitive. For many of the combinations of the confounder correlations, the indirect effect results for the correlated augmented model remained the same as the ones for the posited model when the no omitted confounder was assumed. However, for the combinations of confounder correlations in medium to large range, the model showed high sensitivity that resulted in an overall lack of model convergence. As a result, we were not able to estimate the indirect effects for medium to large confounder correlations.

CONCLUSION

A critical, yet untestable assumption in mediation analysis is the no omitted confounder assumption. This assumption states that an omitted confounder should not influence any pair of variables in a mediation model. Even when the antecedent variable (X) is randomized, one cannot rule out the effect of a confounder on the relationship between the mediator and outcome variable because the values of mediator (M) are not randomized. A more complicated situation is when the antecedent variable is not randomized and when we have two covarying mediators. For this model, a confounder could affect any pair of variables including the antecedent variable. Because the no omitted confounder assumption is untestable, researchers recommend conducting sensitivity analysis that ascertain the impact of potential confounders on the estimates and the possible inference about indirect effects (VanderWeele and Arah, 2011; Tofighi et al., 2013, 2019; Tofighi and Kelley, 2016; Valente et al., 2017). In this manuscript, we extend sensitivity analysis to a nonrandomized latent growth curve mediation model (LGCMM) with two covarying mediators in SEM framework. Conducting sensitivity analysis for the nonrandomized LGCMM has not been done before because certain challenges have interfered. First, nonrandomization means a confounder can impact the antecedent as well as the mediators and the outcome variable. A confounder may impact not only the relationships between the mediators and the outcome variable (as in a randomized mediation model) but also may affect additional relationships of the antecedent to each mediator variable and to the outcome variable. Second, a longitudinal model requires a more sophisticated statistical technique such as LGCMM that can address dependency in repeated measures while modeling mediation through two latent variables: latent intercept and slope. In LGCMM, when the is no covariate or the covariates do not affect the antecedent variable, the antecedent variable is exogenous. The issue remains on how to model and estimate biasing impact of a confounder on the exogenous antecedent variable in LGCMM. Further, the existence of two covarying mediators requires that the indirect effect through each mediator be simultaneously estimated. Conducting sensitivity analysis for each mediator separately while ignoring the other covarying mediators, as is done in a single-mediator model, is likely to result in bias because the two mediators are covarying (VanderWeele, 2015). Thus, techniques developed for a single-mediator model cannot be directly used to conduct sensitivity analysis in a two covarying mediator model. Lastly, given a variety of patterns of confounding bias, summarizing the impact of confounding bias succinctly enables researchers to assess sensitivity of the parameter estimate as well as statistical inference to the confounding bias.

We extended the sensitivity analysis technique termed CAMSA to a nonrandomized LGCMM. A major contribution of our method is the extension of sensitivity analysis to a nonrandomized antecedent variable. This expansion is significant because nonrandomized studies are common and pose additional challenges such as having to address more confounder relationships between the variables because of not having a randomized antecedent variable. Another contribution of our model is that we analytically showed that CAMSA is statistically equivalent to a model with an augmented latent confounder. The analytic work is important because we show how confounder correlations in CAMSA are directly a function of confounder effects in the equivalent latent augmented model. Without explicitly showing these relationships, what confounder correlation is modeling in CAMSA is unclear. Another advantage of our proposed method is that it is performed in SEM framework. The SEM framework allows researchers to, first, simultaneously estimate indirect effects through covarying mediators. Estimating indirect effects independently using separate regression equations would result in biased estimates of indirect effects. Second, researchers could check the effect of confounder correlations on model (global) fit and convergence of correlated augmented model in addition to modeling confounder effects on (local fit of) indirect effects. Examining convergence of the mediation model is a strength of using SEM to conduct mediation analysis and CAMSA because of the simultaneous estimation of multiple regression equations allows researchers to examine the convergence and the fit of mediation models to the sample data. If a specific range of confounder correlations could result in nonconvergence, then checking confounder effect on indirect effect would not be feasible. Third, as shown in the empirical example, existing SEM software can be used to conduct our proposed CAMSA. We provided code in R that would facilitate researchers in conducting the proposed CAMSA in their own research.

We recommend that researchers conduct sensitivity analysis and report the results to assess the robustness of mediation analysis to untestable assumption of no omitted confounders. Because the researchers in social science often use SEM to conduct a mediation analysis, our proposed method along with provided code for CAMSA in SEM context should be an attractive tool that would help researchers enhance robustness of their findings. In addition, we recommend researchers report, in a meaningful way, the range of values for which the results for indirect effects change; that is, investigators should describe when the inference about indirect effect changes. Finally, given the strength of SEM in assessing model fit and convergence of a model, we recommend researchers report ranges of confounder correlations that would result in nonconvergence. We further encourage researchers to explore the reasons for nonconvergence. One reason for nonconvergence is that the correlated augmented model is either inaccurately specified or too constrained to be supported by the sample data. That is, the specific range of confounder correlations render a model not supported by the data. For example, if zero to small range of confounder correlation would cause a large percentage of the nonconvergence, then one might conclude that the posited mediation model is itself sensitive and not robust to small changes. That conclusion would call into question the correct specification of the posited model and could motivate researchers to examine and modify the model carefully. If the model convergence is sensitive to medium and large range of confounder correlations, then researchers could reexamine specification of the posited mediation model. If the convergence rate would not improve, then the researchers could conclude that the model is robust to small range of confounder correlation but not to the medium and large values. The implication of each conclusion should be interpreted in the context of substantive research.

Limitations of the current research are that X, M, and Y are all continuous variables and that we, therefore, assume all the relationships are linear. Future research should extend these methods to a mediation model with one or more categorical outcome. Categorical outcome would require using generalized linear mixed model, and, thus, definition of indirect effects in the potential outcome framework should be used (VanderWeele, 2015). A further limitation of the current study is that we assumed that all variables in the model are measured without errors. While the latent intercept and slope can model measurement error, we did not use latent variables for X and Y. Future research should investigate the joint effects of confounder bias and measurement error (Fritz et al., 2016).

In sum, it is critical to conduct sensitivity analysis to ascertain robustness of the mediation analysis and carefully explain mediation analysis results in the context of correlation confounders and substantive research. Our proposed sensitivity analysis provides a tool for researchers to conduct sensitivity analysis for a nonrandomized LGCMM using available SEM software.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Access to COMBINE data files is restricted to the persons with a signed data access agreement from NIAAA. Requests to access these datasets should be directed to Raye Litten of NIAAA, rlitten@mail.nih.gov.

AUTHOR CONTRIBUTIONS

DT contributed to conception, design, and analysis of the study.

FUNDING

This current study was funded by NIAAA R01 AA025539 (Tofighi and Witkiewitz, MPIs).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.755102/full#supplementary-material

REFERENCES

Albert, J. M., and Wang, W. (2015). Sensitivity analyses for parametric causal mediation effect estimation. *Biostatistics* 16, 339–351. doi: 10.1093/biostatistics/kxu048

American Psychiatric Association (1995). *Diagnostic and Statistical Manual of Mental Disorders*, 4th Edn. Washington, DC: American Psychiatric Publishing.

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice: a review and recommended two-step approach. *Psychol. Bull.* 103, 411–423. doi: 10.1037/0033-2909.103.3.411

Bind, M.-A. C., VanderWeele, T. J., Coull, B. A., and Schwartz, J. D. (2016). Causal mediation analysis for longitudinal data with exogenous exposure. *Biostatistics* 17, 122–134. doi: 10.1093/biostatistics/kxv029

Bogoya, J., Böttcher, A., and Grudsky, S. (2012). Eigenvalues of Hermitian Toeplitz matrices with polynomially increasing entries. *J. Spectr. Theory* 2, 267–292. doi: 10.4171/JST/29

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn. Mahwah NJ: Erlbaum.

Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., and MacKinnon, D. P. (2013). Sensitivity plots for confounder bias in the single mediator model. *Eval. Rev.* 37, 405–431. doi: 10.1177/0193841X14524576

Daniel, R. M., De Stavola, B. L., Cousens, S. N., and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics* 71, 1–14. doi: 10.1111/biom.12248

Derogatis, L. R., and Melisaratos, N. (1983). The brief symptom inventory: an introductory report. *Psychol. Med.* 13, 595–605. doi: 10.1017/S0033291700048017

DiClemente, C. C., Carbonari, J. P., Montgomery, R. P., and Hughes, S. O. (1994). The alcohol abstinence self-efficacy scale. *J. Stud. Alcohol* 55, 141–148.

DiClemente, C. C., and Hughes, S. O. (1990). Stages of change profiles in outpatient alcoholism treatment. *J. Subst. Abus.* 2, 217–235. doi: 10.1016/S0899-3289(05)80057-4

Fritz, M. S., Kenny, D. A., and MacKinnon, D. P. (2016). The combined effects of measurement error and omitting confounders in the single-mediator model. *Multivariate Behav. Res.* 51, 681–697. doi: 10.1080/00273171.2016.1224154

Harring, J. R., McNeish, D. M., and Hancock, G. R. (2017). Using phantom variables in structural equation modeling to assess model sensitivity to external misspecification. *Psychol. Methods* 22, 616–631. doi: 10.1037/met0000103

Hartzler, B., Witkiewitz, K., Villarroel, N., and Donovan, D. (2011). Self-efficacy change as a mediator of associations between therapeutic bond and one-year outcomes in treatments for alcohol dependence. *Psychol. Addict. Behav.* 25, 269–278. doi: 10.1037/a0022869

Higham, N. J. (2002). Computing the nearest correlation matrix—A problem from finance. *IMA J. Numer. Anal.* 22, 329–343. doi: 10.1093/imanum/22.3.329

Hong, G., Qin, X., and Yang, F. (2018). Weighting-based sensitivity analysis in causal mediation studies. *J. Educ. Behav. Stat.* 43, 32–56. doi: 10.3102/1076998617749561

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Stat. Sci.* 25, 51–71. doi: 10.1214/10-STS321

Imai, K., and Yamamoto, T. (2013). Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Polit. Anal.* 21, 141–171.

Judd, C. M., and Kenny, D. A. (1981). Process analysis. *Eval. Rev.* 5, 602–619. doi: 10.1177/0193841X8100500502

Lindmark, A., de Luna, X., and Eriksson, M. (2018). Sensitivity analysis for unobserved confounding of direct and indirect effects using uncertainty intervals. *Stat. Med.* 37, 1744–1762. doi: 10.1002/sim.7620

MacKinnon, D. P., and Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Pers. Soc. Psychol. Rev.* 19, 30–43. doi: 10.1177/1088868314542878

Maisto, S. A., Roos, C. R., O'Sickey, A. J., Kirouac, M., Connors, G. J., Tonigan, J. S., et al. (2015). The indirect effect of the therapeutic alliance and alcohol abstinence self-efficacy on alcohol use and alcohol-related problems in project MATCH. *Alcoholism* 39, 504–513. doi: 10.1111/acer.12649

McCandless, L. C., and Somers, J. M. (2019). Bayesian sensitivity analysis for unmeasured confounding in causal mediation analysis. *Stat. Methods Med. Res.* 28, 515–531. doi: 10.1177/0962280217729844

Miller, W. R. (1996). *Form 90: A structured Assessment Interview for Drinking and Related Behaviors*. New Delhi: National Institute on Alcohol Abuse and Alcoholism.

Moyers, T. B., Martin, T., Houck, J. M., Christopher, P. J., and Tonigan, J. S. (2009). From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *J. Consult. Clin. Psychol.* 77, 1113–1124. doi: 10.1037/a0017189

Pearl, J. (2014). Interpretation and identification of causal mediation. *Psychol. Methods* 19, 459–481. doi: 10.1037/a0036434

Podsakoff, P. M., MacKenzie, S. B., 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

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing* (**??**) *[Computer Software]*. Vienna: R Foundation for Statistical Computing.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edn. Thousand Oaks, CA: SAGE.

Robins, J. M., and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 3, 143–155.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36.

Rousseeuw, P. J., and Molenberghs, G. (1994). The shape of correlation matrices. *Am. Stat.* 48, 276–279. doi: 10.2307/2684832

Schott, J. R. (1997). *Matrix Analysis for Statistics*. Hoboken, NJ: Wiley.

Singer, J. D., and Willett, J. B. (2003). *Applied Longitudinal Data Analysis Modeling Change and Event Occurrence*. Oxford: Oxford University Press.

Skinner, H. A., and Allen, B. A. (1982). Alcohol dependence syndrome: measurement and validation. *J. Abnorm. Psychol.* 91, 199–209. doi: 10.1037/0021-843X.91.3.199

Snijders, T. A. B., and Bosker, R. J. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, 2nd Edn. Thousand Oaks, CA: SAGE.

Talloen, W., Moerkerke, B., Loeys, T., De Naeghel, J., Van Keer, H., and Vansteelandt, S. (2016). Estimation of indirect effects in the presence of unmeasured confounding for the mediator–outcome relationship in a multilevel 2-1-1 mediation model. *J. Educ. Behav. Stat.* 41, 359–391. doi: 10.3102/1076998616636855

The COMBINE Study Research Group (2003). Testing combined pharmacotherapies and behavioral interventions for alcohol dependence (The COMBINE Study): a pilot feasibility study. *Alcoholism* 27, 1123–1131. doi: 10.1097/01.ALC.0000078020.92938.0B

Tofighi, D., Hsiao, Y.-Y., Kruger, E. S., MacKinnon, D. P., Van Horn, M. L., and Witkiewitz, K. (2019). Sensitivity analysis of the no-omitted confounder assumption in latent growth curve mediation models. *Struct. Equ. Modeling* 26, 94–109. doi: 10.1080/10705511.2018.1506925

Tofighi, D., and Kelley, K. (2016). Assessing omitted confounder bias in multilevel mediation models. *Multivariate Behav. Res.* 51, 86–105. doi: 10.1080/00273171.2015.1105736

Tofighi, D., West, S. G., and MacKinnon, D. P. (2013). Multilevel mediation analysis: the effects of omitted variables in the 1-1-1 model. *Br. J. Math. Stat. Psychol.* 66, 290–307. doi: 10.1111/j.2044-8317.2012.02051.x

Valente, M. J., Pelham, W. E. I., Smyth, H., and MacKinnon, D. P. (2017). Confounding in statistical mediation analysis: what it is and how to address it. *J. Couns. Psychol.* 64, 659–671. doi: 10.1037/cou0000242

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.

VanderWeele, T. J., and Arah, O. A. (2011). Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* 22, 42–52. doi: 10.1097/EDE.0b013e3181f74493

von Soest, T., and Hagtvet, K. A. (2011). Mediation analysis in a latent growth curve modeling framework. *Struct. Equ. Modeling* 18, 289–314. doi: 10.1080/10705511.2011.557344

Ware, J. Jr., Kosinski, M., and Keller, S. D. (1996). A 12-item short-form health survey: construction of scales and preliminary tests of reliability and validity. *Med. Care* 34, 220–233. doi: 10.1097/00005650-199603000-00003

Wicklin, R. (2015). *A simple way to construct a large correlation matrix. The DO Loop*. Available online at: https://blogs.sas.com/content/iml/2015/09/23/large-spd-matrix.html (accessed September 23, 2015).

World Health Organization (1997). *WHOQOL: Measuring Quality of Life*. Geneva: World Health Organization.

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

SYSTEMATIC REVIEW
published: 07 December 2021
doi: 10.3389/fpsyg.2021.764209

Reliability, and Convergent and Discriminant Validity of Gaming Disorder Scales: A Meta-Analysis

*Seowon Yoon[1], Yeji Yang[1], Eunbin Ro[1], Woo-Young Ahn[2], Jueun Kim[3], Suk-Ho Shin[4], Jeanyung Chey[2] and Kee-Hong Choi[1]**

[1] School of Psychology, Korea University, Seoul, South Korea, [2] Department of Psychology, Seoul National University, Seoul, South Korea, [3] Department of Psychology, Chungnam National University, Daejeon, South Korea, [4] Dr. Shin's Neuropsychiatric Clinic, Seoul, South Korea

Background: An association between gaming disorder (GD) and the symptoms of common mental disorders is unraveled yet. In this preregistered study, we quantitatively synthesized reliability, convergent and discriminant validity of GD scales to examine association between GD and other constructs.

Methods: Five representative GD instruments (GAS-7, AICA, IGDT-10, Lemmens IGD-9, and IGDS9-SF) were chosen based on recommendations by the previous systematic review study to conduct correlation meta-analyses and reliability generalization. A systematic literature search was conducted through Pubmed, Proquest, Embase, and Google Scholar to identify studies that reported information on either reliability or correlation with related variables. 2,124 studies were full-text assessed as of October 2020, and 184 were quantitatively synthesized. Conventional Hedges two-level meta-analytic method was utilized.

Results: The result of reliability generalization reported a mean coefficient alpha of 0.86 (95% CI = 0.85–0.87) and a mean test-retest estimate of 0.86 (95% CI = 0.81–0.89). Estimated effect sizes of correlation between GD and the variables were as follows: 0.33 with depression (k = 45; number of effect sizes), 0.29 with anxiety (k = 37), 0.30 with aggression (k = 19), –0.22 with quality of life (k = 18), 0.29 with loneliness (k = 18), 0.56 with internet addiction (k = 20), and 0.40 with game playtime (k = 53), respectively. The result of moderator analyses, funnel and forest plots, and publication bias analyses were also presented.

Discussion and Conclusion: All five GD instruments have good internal consistency and test-retest reliability. Relatively few studies reported the test-retest reliability. The result of correlation meta-analysis revealed that GD scores were only moderately associated with game playtime. Common psychological problems such as depression and anxiety were found to have a slightly smaller association with GD than the gaming behavior. GD scores were strongly correlated with internet addiction. Further studies should adopt a rigorous methodological procedure to unravel the bidirectional relationship between GD and other psychopathologies.

OPEN ACCESS

Edited by:
Marta Martín-Carbonell, Universidad Cooperativa de Colombia, Colombia

Reviewed by:
Sai-fu Fung, City University of Hong Kong, Hong Kong SAR, China
Danka Purić, University of Belgrade, Serbia

****Correspondence:***
Kee-Hong Choi
kchoi@korea.ac.kr

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

Received: *25 August 2021*
Accepted: *10 November 2021*
Published: *07 December 2021*

Citation:
Yoon S, Yang Y, Ro E, Ahn W-Y, Kim J, Shin S-H, Chey J and Choi K-H (2021) Reliability, and Convergent and Discriminant Validity of Gaming Disorder Scales: A Meta-Analysis. Front. Psychol. 12:764209. doi: 10.3389/fpsyg.2021.764209

Limitations: The current study did not include gray literature. The representativeness of the five tools included in the current study could be questioned. High heterogeneity is another limitation of the study.

Systematic Review Registration: [https://www.crd.york.ac.uk/PROSPERO/], identifier [CRD42020219781].

Keywords: gaming disorder (GD), meta-analysis, convergent validity, discriminant validity, reliability generalization meta-analysis, validity generalization, association

INTRODUCTION

Since games are one of the most popular leisure activities worldwide, they are now available almost everywhere via computers, mobile phones, and tablets. Generally, gamers enjoy gaming as a leisure activity, and the games seem to affect them positively (Jones et al., 2014). Increasing concerns, however, have been raised about excessive gaming behaviors. American Psychiatric Association (2013) has already introduced the provisional diagnostic criteria for internet gaming disorder in Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5). The World Health Organization (WHO) recently adopted gaming disorder (GD) as a diagnosis in the eleventh edition of the International Classification of Diseases (World Health Organization [WHO], 2018). Despite the few discrepancies in the diagnostic criteria for GD in ICD-11 and DSM-5, the common symptoms of GD include continuation of gaming and impaired control over gaming behavior, which result in functional impairments (Jo et al., 2019).

The official listing of GD diagnosis is debatable (Aarseth et al., 2017; Griffiths et al., 2017; Király and Demetrovics, 2017; Kuss et al., 2017; Van Den Brink, 2017; Rumpf et al., 2018; Van Rooij et al., 2018). Several high-quality studies including epidemiological studies (Lemmens et al., 2015; Pontes et al., 2016; Wittek et al., 2016; Han et al., 2018), clinical outcome studies (see King et al., 2017), neuroimaging studies (Fauth-Bühler and Mann, 2017; Han et al., 2017; Liu et al., 2018), and experimental studies (Sariyska et al., 2017; Kräplin et al., 2021) have been published in the recent years, showing improvements with regard to the quality of studies and methodological issues raised by researchers (Petry and O'Brien, 2013; Van Rooij et al., 2018). Most studies, nonetheless, have relied on self-report assessment tools rather than relying on structured clinical interviews, which is partially due to the inconsistency in definition and the different diagnostic criteria (Jeong et al., 2018). Whether the assessment tools are reliable and whether they could validly measure GD are important questions that should be answered.

Another unresolved but important issue is the association between GD and the symptoms of common mental disorders (see Billieux et al., 2017; Van Rooij et al., 2018). Pontes and Griffiths (2019) commented the importance of key risk factors related to comorbidities. Literature has reported mixed results in the association between gaming disorder and psychiatric disorders. Associations between gaming disorder and the common symptoms of mental disorders were found to be considerably weaker than between symptoms of other disorders at least in young age group (Wichstrøm et al., 2018). In contrast, some studies have reported that the underlying mental illness can be a strong predictor of problematic gaming (Kardefelt-Winther, 2014; Billieux et al., 2015), perhaps even a cause (Van Rooij et al., 2018). Authors also have different interpretations for the association. Some authors consider strong association between GD and mental disorders a natural result because clinicians seldomly assess GD without considering comorbidities (Wichstrøm et al., 2019). On the other hand, strong association is also a basis for supporting the idea that GD may be a consequence of other mental disorders (Van Rooij et al., 2018).

In the current study, we focused on construct validity among several aspects of validity since convergent and discriminant validity provide information on the association between GD and other constructs. Reliability and construct validity provide information on what GD assessment tools consistently measure. Poor construct validity of the measure limits the ability of the tools to achieve its intended purpose of measurement because it remains unclear whether the GD instruments represent the construct of the GD or other psychopathological features. If GD instruments have enough construct validity, the association between GD and gaming behavior would be expected to have stronger association compared to the associations between GD and other psychopathological variables.

To our knowledge, no study has systematically examined association between GD scales and symptoms of common psychiatric comorbidities and compared it to the association between GD and gaming behavior. The recent studies on psychological science adopted the reliability generalization and the correlation meta-analytic technique to perform a meta-analysis of a sample of studies with the purpose of estimating the population reliability and population correlation value of the respective studies (Rodriguez and Maeda, 2006; López-Pina et al., 2015; Miller et al., 2018). In the current study, we quantitatively synthesized the bivariate Pearson's correlation coefficients between GD assessment tools and common psychological problem (e.g., depression, anxiety, aggression) scales, which refers to the statistic of construct validity, to examine the association between GD and psychological variables. We also conducted reliability generalization to examine the consistency of the scales.

Recently, King et al. (2020) reviewed 32 GD assessment tools in their qualitative review paper, recommending five GD instruments with relatively great evidential support. The five tools are 7-item Game Addiction Scale (GAS-7; Lemmens et al., 2009), 9-item Internet Gaming Disorder Scale-Short Form (IGDS9-SF;

Pontes and Griffiths, 2015), 10-item Internet Gaming Disorder Test (IGDT-10; Király et al., 2017), Assessment of Internet and Computer Addiction Scale-Gaming (AICA; Müller et al., 2014), and Lemmens Internet Gaming Disorder Scale-9 (Lemmens IGD-9; Lemmens et al., 2015). Among excluded instruments, Young Internet Addiction scale (Young, 1998) is the most frequently utilized scale, and Young Diagnostic Questionnaire (Young, 1998) is the most cited instrument (King et al., 2020). However, they are relatively old scales and are more related to internet addiction rather than GD. In general, YIAT, GAS-7, and IGDS9-SF are frequently used in the field, and IGDT-10 is an instrument that is evenly used in both the West and the East (King et al., 2020). King et al. (2020) recommended the five tools in consideration of the following factors: DSM -5 and ICD-11 coverage, existence of longitudinal studies, adaptation of structured interview, validation of reliability and cut-off score, dimensionality, criterion validity, test refinement and impairment. Divergent validity, however, was not examined by King et al. (2020). Given the importance of the association between GD and other mental disorders, synthesizing and comparing the magnitude of convergent and discriminant validity can significantly contribute to the understanding of GD.

The GD studies often operationalized the convergent validity as there is a bivariate association between a gaming behavior (i.e., hours per week spent gaming) and a score on a GD tool (King et al., 2020). The given association between a score on a GD tool and a gaming behavior represents convergent validity. The associations between the GD tools and other variables can be operationalized as discriminant validity. In a recent article of theirs, Rönkkö and Cho (2020) provided a general definition of discriminant validity. A discriminant validity means that the two measures intended to measure distinct constructs have discriminant validity if the absolute value of the correlation between the measures after correcting for measurement error is low enough for the measures to be regarded as measuring distinct constructs (Rönkkö and Cho, 2020). If the associations between GD and other psychological variables are too strong, the GD tools may reveal the weaknesses in discriminant validity and present the diagnostic needs from the other psychiatric disorders. If the associations are too small, it might not properly reflect the pain and burden of problematic gaming. By quantitatively synthesizing the correlation coefficients to estimate convergent and discriminant validity coefficient, we can quantify and compare the magnitude of each association between GD and other variables.

This study's objectives are to (1) synthesize the reliability coefficients; (2) examine the convergent and discriminant validity of the GD tools, further investigating the overall association between the GD tools and other psychological/behavioral variables; and (3) investigate how the study characteristics and potential moderator variables affect the reliability and validity estimates, wherein the potential influencing variables include the specific GD instrument used in the study, the type of the sample, study location, and gender ratio of the study participants. Demographic variables such as age, gender, and study location are variables often examined for measurement invariance in this field (see King et al., 2020), and significant moderators of the prevalence rate of GD (see Andreetta et al., 2020; Stevens et al., 2021). Since five scales which cover different domain of diagnostic criteria were included, we did not perform quantitative synthesis on factor structure in order to prevent confusion. Since there is no gold standard for GD diagnosis, and only few studies adopted rigorous clinical interview, we were unable to conduct a meta-analysis for predictive validity of GD assessment tools.

METHODS

Search Strategy

The current study was conducted based on the PRISMA statement (Moher et al., 2009; Page et al., 2021) and recommendations received for the correlational meta-analyses (Quintana, 2015). PRISMA checklist (Page et al., 2021) is included in **Supplementary Material 1**. The protocol for the current study has been preregistered on PROSPERO (CRD42020219781). While full electronic search strategy for databases using search terms is a standard procedure for the systematic review, the search strategy in the current study was modified because too many irrelevant and unqualified studies were searched with broad search terms, whereas too many missing studies were searched when narrowing the scope. The first database search for all the published studies with GD assessment tools was executed in PubMed, Proquest, and Embase on August 18, 2020, resulting in 1,343 potentially eligible articles. However, we found too many relevant studies were missing. Great heterogeneity in articles of diagnostic criteria (e.g., DSM-5 and ICD-11 from WHO), type of gaming (e.g., mobile, computer, video-only, online, smartphone gaming), name of the disorder and key-terms (e.g., game addiction, internet gaming, online gaming, video gaming, problematic, overuse, excessive) were factors that made standard search procedure ineffective.

Therefore, we modified our search strategy by selecting a few GD scales to be included in advance. Since King et al. (2020) nicely reviewed 32 GD assessment tools in qualitative way, the five recommended tools with great evidential support were chosen to extract and synthesize the correlation data. The second database search included all the empirical studies that had employed at least one of the 5 GD assessment tools. The search was carried out via two different procedures: (1) A computer-based search of Pubmed, Proquest, and Embase using broad keywords to ensure that all studies adopted one of the five scales are included (e.g., IGDS AND (SF OR short OR 9) not to omit any empirical studies that adopted IGDS9-SF), and (2) a procedural collection of the all Google Scholar citation records for the five tools (as of October 2020). The duplicates of the identified articles were first eliminated by using the Endnote software[1] version 20 followed by double-checking from the authors. Search strategy of the current study is provided in the **Supplementary Material 2**.

Inclusion and Exclusion Criteria

Articles were included if they (a) were peer-reviewed journal articles, (b) used one of the five tools recommended by the

[1]http://endnote.com

current systematic review paper, (c) reported the reliability coefficient or bivariate correlation coefficient via the scales of depression, anxiety, aggression, loneliness, quality of life, internet addiction and game playtime, and (d) written in English. Articles were excluded if they (a) did not include relevant information for GD, (b) non-empirical studies such as meta-analyses and systematic review papers, or (c) did not include the reliability or validity coefficient. Due to difficulties in searching, data extracting, and assessing the study quality, we decided to include the articles which were published with the peer-reviewed process.

Coding Procedure

All the preselected variables were coded. The coded variables included demographic information of the study, name of the utilized assessment tool, psychometric information, and bivariate Pearson's correlation coefficient. The potentially eligible articles were systematically coded by three co-authors, namely, SY, YY, and ER. For the longitudinal studies that reported repetitive information using the same sample, we coded the information reported during the first wave. This is because it often contains a larger sample that that during the second or third wave. For multisite cross-sectional studies that included more than one effect size, information for the rest of effect sizes were coded separately. For studies that used various scales to measure only one psychological variable, the effect sizes were integrated into one effect size by calculating the average.

The candidate studies for data synthesis were evenly split between three raters SY, YY, and ER, and then cross-checked by the corresponding author independently. Overall, the level of agreement on the coding was 92.7%, and all the coded information was reached an agreement. A copy of the coding sheet is available in the **Supplementary Material 3**.

Selection Process

After the elimination of duplicates using Endnote software, 605 articles were identified via the database keyword search and 1,519 articles were identified via the Google Scholar citation records. Total of 2,124 studies were full-text screened to identify the potentially eligible studies based on the inclusion and exclusion criteria. We found and removed duplicates within and between each database. There were 135 overlapping studies within the Google scholar citation records, and 37 overlapping studies between electronic database search records and Google scholar citation records. As a result, 249 potentially eligible studies were identified. E-mails requesting additional data were sent to the corresponding authors of 49 studies. As of February 2021, 17 authors (34.7%) had responded to the request, and the information provided was finally included in the quantitative synthesis. As a result, 184 of the 249 studies were quantitatively synthesized and 65 were excluded. Among 65 excluded studies, 33 did not include any information on the variables of our interest. The rest 32 studies were excluded due to no reply to the inquiry. **Figure 1** presents a flowchart of the database search, screening, and data coding process. The list of the included studies is provided in **Supplementary Material 4**.

Meta-Analytic Method and Statistical Analysis for Reliability Generalization

Reliability generalization is a powerful tool to characterize the mean measurement error variance across studies, and also the variabilities in score quality and the study features (Vacha-Haase, 1998). We utilized this technique to estimate the overall level of reliability of the included studies and to find differences in the level of reliability among the five instruments. Separate meta-analyses were conducted for reliability generalization and validity generalization. The current study utilized a meta-analytic technique to quantitatively synthesize the findings of various studies and examine the overall reliability of the GD assessment tools that are frequently used. Cronbach's alpha coefficients (Cronbach, 1951) were frequently reported, allowing us to synthesize the findings. Information on test-retest reliability, however, was less frequently reported. To conduct reliability generalization of the internal consistency, we extracted the Cronbach's alpha coefficient for just the total score of the five GD assessment tools drawn from eligible studies. Cronbach's alpha coefficient offers information on the internal consistency of the test scale (Tavakol and Dennick, 2011). With regard to the calculation of the mean coefficient alpha, Bonett's (2002) transformation was applied to normalize the distribution and stabilize their variance: $Li = Ln(1-\alpha i)$, where Ln is the natural logarithm. After synthesizing reliability with transformed values, we converted the Bonett-transformed metric back to the original metric of Cronbach's alpha coefficient to facilitate interpretation. The test-retest reliability coefficients reported from the included studies were descriptively presented in the result. We adopted the same correlation meta-analysis technique for the quantitative synthesis of test-retest reliability coefficients since test-retest reliability is often measured with a correlation coefficient.

Meta-Analytic Method and Statistical Analysis for Validity Generalization

We coded all bivariate correlation coefficients between GD and psychological variables if the number of effect sizes is sufficient enough to conduct quantitative synthesis ($j > 10$). We considered the correlation between a GD scale score and the game playtime as a convergent validity variable. Depression, anxiety, impulsivity, loneliness, aggression, gambling addiction, internet addiction, alcohol addiction, and quality of life (QOL) were considered potential discriminant validity variables. Among ten variables, gambling addiction ($j = 5$, number of effect sizes), alcohol addiction ($j = 2$), and impulsivity ($j = 6$) were excluded due to the insufficient number of effect sizes for quantitative synthesis. As a result, we performed quantitative synthesis of correlation between GD and seven psychological variables: depression, anxiety, quality of life (QOL), aggression, loneliness, internet addiction, and game playtime.

To estimate the overall mean effect size and correlation coefficient, the current literature has dominantly adopted two approaches (Field and Gillett, 2010; Brannick et al., 2019). These two approaches were proposed by Schmidt and Hunter (1998) and Hedges (Hedges, 1992; Hunter and Schmidt, 2004; Borenstein et al., 2011). However, determining which approach

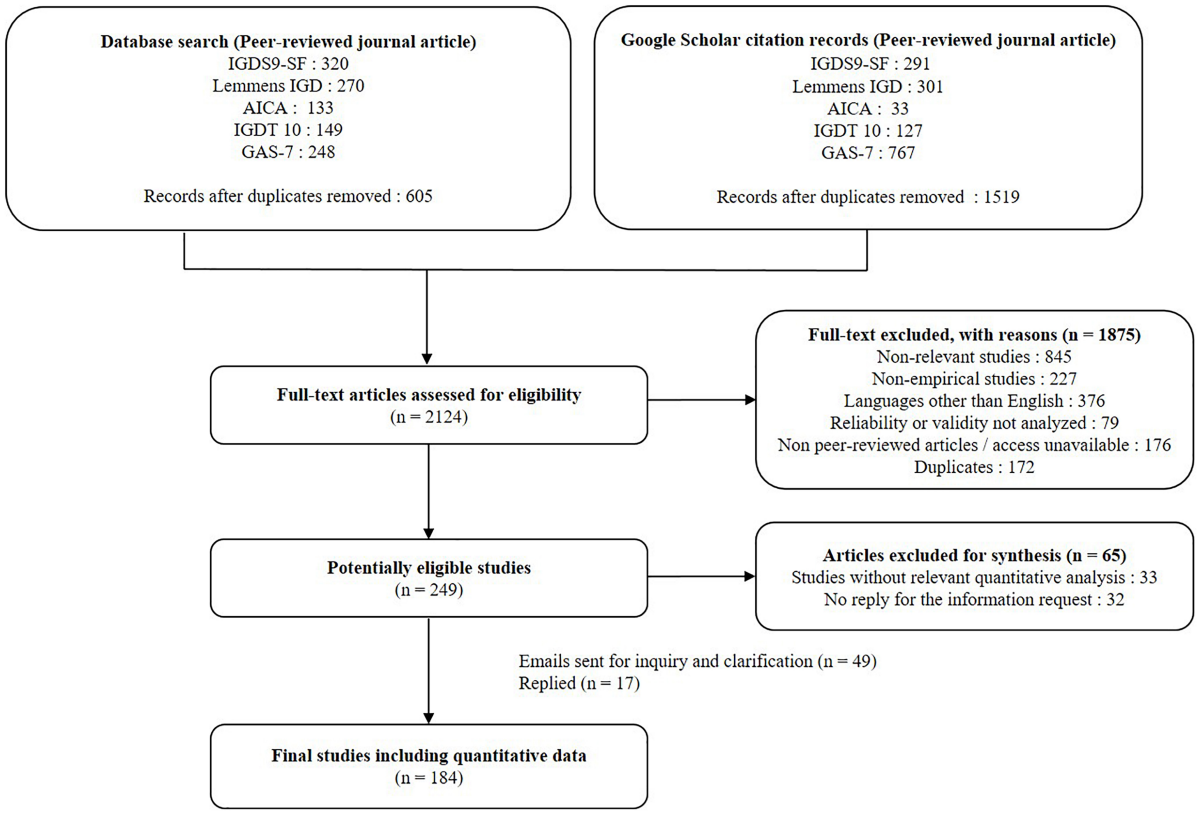

FIGURE 1 | Flow chart of the search process in accordance with PRISMA guidelines.

is more appropriate for the correlation coefficient's meta-analysis has been controversial (Field, 2005; Field and Gillett, 2010). In addition to the two commonly adopted techniques, Brannick et al. (2019) also introduced a novel estimator, providing better coverage and slightly better credibility values than the commonly used approaches. These meta-analytic methods are based on the random-effects model. A random-effects model allows the true effect to differ in each study, whereas a fixed effect model assumes all the studies share a common effect size (Borenstein et al., 2010). As the studies included in this meta-analysis were conducted in different regions and have differ samples, a random effects model was used to derive the effect size and confidence level.

For correcting measurement unreliability, Hunter-Schimidt estimator (Hunter and Schmidt, 2004). Morris estimator (Brannick et al., 2019) apply the individual correction technique to estimate the mean effect size. Hedges method (Borenstein et al., 2011), however, does not adopt the individual correction technique to estimate the effect size. As the current study also aims to conduct reliability generalization to examine the reliability of the GD assessment tools, we utilized the Hedges method. The current study adopted a conventional two-level meta-analytic method instead of a three-level model or robust variance estimation technique to estimate the pooled effect size of the correlation. Although a three-level model and robust variance estimation technique have several advantages over a conventional two-level meta-analytic model (Hedges et al., 2010; Assink and Wibbelink, 2016; Harrer et al., 2021), scarce information on the variance of effect size within individual studies made it difficult to apply a three-level model or robust variance estimation method. We therefore conducted the conventional two-level meta-analysis in the current study.

Heterogeneity and Moderator Analyses

As the current study synthesized the findings of studies that used five different assessment tools, a high heterogeneity was expected. To examine the heterogeneity of the quantitative synthesis, the current study reported Tau (T), Tau-squared (T^2), and I^2 as the measures of heterogeneity between the studies. Tau and Tau-squared are reported in the same metric as the effect size, providing information about the dispersion of true effects on the absolute scale (Borenstein et al., 2017). A guide to interpret the I^2 statistic (Borenstein et al., 2017) is as follows: small heterogeneity (I^2 25%), moderate heterogeneity (I^2 50%), and considerable heterogeneity (I^2 = 75%).

Categorical moderator analyses were conducted to identify the potential impacts of reliability and validity generalizations. One study characteristic moderator, (a) the specific GD instrument used in the study (categorized into "IGDS9-SF," "GAS-7," "Lemmens IGD-9," "AICA," and "IGDT-10"), was considered the potential impact for reliability generalization. Three study characteristics were considered as the potential categorical moderators for validity generalization, namely, (a) the specific

GD instrument used in the study, (b) the type of the sample (categorized into "adolescents," "adults," and "both"), and (c) the study location (categorized into six continents). Categorical moderator analyses were conducted when each of the subgroups had at least 4 studies. Fu et al. (2011) suggested that each subgroup should have at least four studies for a categorical moderator analysis. Some subgroups with an insufficient number of studies (less than four studies) were excluded from the moderator analysis. To investigate whether the continuous moderator (d) gender ratio affects effect sizes, we performed a meta-regression with the ratio of male participants.

Statistical Software

The statistical analysis was conducted in R software (version 4.0.3) using metafor (Viechtbauer, 2010), meta (Schwarzer, 2007), and dmetar packages (Harrer et al., 2019). The packages provide various functions to facilitate study synthesis. These include moderator analysis, meta-regression analysis, Egger's regression test (Egger et al., 1997), sensitivity meta-analysis for publication bias, and various types of meta-analytical plotting.

Publication Bias

Rothstein et al. (2005) suggested that publication bias, also known as file-drawer problem, could occur since studies without statistically significant results are less likely to be published. The current study examined the risk of publication bias by drawing a funnel plot and conducting Egger's test (Egger et al., 1997). Egger's regression test quantifies the funnel plot asymmetry and performs a statistical test. If the *p*-value of Egger's test is significant, the significant asymmetry in the Funnel plot caused by the publication bias or "small study effects" is indicated (Sterne et al., 2001). Cumulative meta-analysis and sensitivity analysis were additionally conducted when Egger's test indicated the presence of publication bias.

RESULTS

Description of Included Studies

The current study included 184 articles that reported the results from 205 independent samples with 285,752 participants. The estimated mean age of the study samples based on the studies' reported statistic was 22.12, and 60.7% of the participants were male. Of the studies included up to December 2020, 159 studies (86.4%) have been published since 2016 and 102 studies (55.4%) since 2019. While 94 studies were conducted in Europe, 61 studies were conducted in Asia. Regarding the targeted age group, 63 studies targeted adult samples, 56 studies targeted adolescent samples, and the remaining 65 included both adult and adolescent samples. Of the 184 studies, 49 conducted factor analysis and reported related statistics. While most of the studies (k = 42) conducted confirmatory factor analysis, two studies conducted exploratory factor analysis and five studies conducted both. IGDS9-SF was found to be the most frequently utilized tool (k = 81, 44.0%). Key characteristics of the included studies are reported in **Table 1**.

TABLE 1 | Key characteristics of the included studies for quantitative synthesis.

Sample size	**n (%)**
Total	285,752 (100%)
Male	173,570 (60.7%)
Female	112,086 (39.2%)
Unknown	97 (0.0%)
Characteristics of studies	**j (%)**
Total	184 (100%)
Sample target	
Adults	63 (34.2%)
Adolescents	56 (30.4%)
Both	65 (35.3%)
Location	
Europe	94 (51.1%)
Asia	61 (33.2%)
North America	10 (5.4%)
South America	1 (0.5%)
Australia/New Zealand	9 (4.9%)
Africa	1 (0.5%)
Global	8 (4.3%)
GD tools	
IGDS9-SF	81 (44.0%)
GAS-7	58 (31.5%)
Lemmens IGD-9	18 (9.8%)
IGDT-10	17 (9.2%)
AICA	10 (5.4%)

n, number of samples; j, number of studies.

Reliability Generalization

Result of Reliability Generalization

Cronbach's alpha coefficient of 193 effect sizes (from 172 studies) were quantitatively synthesized for the respective reliability generalization. The number of studies reporting the Cronbach's alpha coefficients of GD assessment tools were as follows: 90 effect sizes from the 76 studies for IGDS9-SF, 58 effect sizes from the 53 studies for GAS-7, 20 effect sizes from the 18 studies for Lemmens IGD-9, 16 effect sizes from the 16 studies for IGDT-10 and, 9 effect sizes from the 9 studies for AICA. All the five assessment tools demonstrated an appropriate level of reliability. The estimated average reliability coefficient obtained from Bonett's transformation was 1.97 (95% CI = 1.90–2.04). Then, to facilitate the interpretation, Bonett's transformed reliability coefficient was transformed back into Cronbach's alpha coefficient. The result of RG reported a mean coefficient alpha of 0.86 (95% CI = 0.85–0.87). The result of RG for each of the five GD Assessment Tools is summarized in **Table 2**. The forest plot of RG is included in the (**Supplementary Figure 1**).

A total of 8 studies reported test-retest reliability ranging from 0.78 to 0.94. The number of studies reporting the test-retest reliability of the GD assessment tools are as follows: 4 studies for IGDS9-SF (0.78–0.94), 3 studies for GAS-7 (0.80–0.83), 1 study for Lemmens IGD-9 (0.83), and none for IGDT-10 and AICA. The estimated pooled coefficient of test-rest reliability was 0.86 (95% CI = 0.81–0.89).

TABLE 2 | Result of reliability statistics for the five GD assessment tools.

						95% CI	80% CR	Heterogeneity		
GD tools	**j**	**k**	**n**	α_{trf}	α	**LL, UL**	**LL, UL**	τ	τ^2	I^2(%)
Total	172	193	263,979	1.97	0.86	[0.85, 0.87]	[0.64, 0.95]	0.48	0.23	99.3
IGDS9-SF	76	90	65,324	2.20	0.89	[0.88, 0.90]	[0.73, 0.95]	0.45	0.20	98.5
GAS-7	53	58	91,132	1.82	0.84	[0.82, 0.85]	[0.65, 0.92]	0.39	0.15	98.9
Lemmens IGD-9	18	20	16,962	1.64	0.81	[0.76, 0.84]	[0.51, 0.92]	0.46	0.21	98.7
IGDT-10	16	16	54,695	1.70	0.82	[0.77, 0.85]	[0.55, 0.93]	0.45	0.20	99.7
AICA	9	9	35,866	1.92	0.85	[0.80, 0.89]	[0.61, 0.95]	0.47	0.23	99.7

j, number of studies; k, number of effect size (Cronbach alpha coefficient of GD tools); n, number of samples; α_{trf}, transformed mean Cronbach's alpha coefficient; α, back-transformed mean Cronbach's alpha coefficient; CI, confidence interval; CR, credibility interval; τ, square root of estimated tau^2; τ^2, estimated amount of total heterogeneity; I^2, total heterogeneity/total variability.

Heterogeneity and Moderator Analysis

The results of the heterogeneity test for reliability were significant for all the included studies ($\tau = 0.483$, $\tau^2 = 0.233$, I^2 =99.3%). To assess the effect of the specific GD instrument used in the study on heterogeneity, a categorical moderator analysis on moderator (a) was conducted. Reliability was revealed to be significantly heterogeneous depending on the measure verified via an omnibus test of hypothesis [QM (4) = 57.56, $p < 0.001$]. Since IGDS9-SF showed the highest Bonett-transformed coefficient alpha, ANOVA was conducted between the measures. All ANOVA comparisons were conducted to examine whether significant difference exists between the magnitude of each coefficient. The results show that the Bonett-transformed coefficient alpha of IGDS9-SF was significantly higher than the coefficients of GAS-7, Lemmens IGD-9, and IGDT-10 (all $p < 0.001$) but was not higher than the coefficient of AICA ($p = 0.06$). The ANOVA result between AICA and Lemmens IGD-9 was also statistically significant ($p < 0.05$).

Publication Bias

Publication bias was assessed using funnel plots and Egger's regression test. The result of Egger's regression test did not indicate the presence of publication bias for IGDS9-SF ($z = 1.37$, $p = 0.17$), Lemmens IGD-9 ($z = -0.76$, $p = 0.45$), IGDT-10 ($z = -0.76$, $p = 0.45$), and AICA ($z = 0.03$, $p = 0.97$). Egger's test of GAS-7, however, indicated the presence of publication bias ($z = -2.02$, $p = 0.04$). Funnel plots are included in the (**Supplementary Figure 2**).

Association and Validity Generalization

Results of Validity Generalization

A total of 210 effect sizes were extracted and synthesized for validity generalization from the 115 studies analyzed. The number of studies reporting the correlation coefficients between GD assessment tools and psychological or behavioral measurement are as follows: 45 effect sizes from the 44 studies for depression, 37 effect sizes from the 36 studies for anxiety, 19 effect sizes from the 17 studies for aggression, 18 effect sizes from the 17 studies for quality of life and loneliness, 20 effect sizes from the 18 studies for internet addiction, and 53 effect sizes from the 51 studies for game playtime. DASS-21(Depression Anxiety Stress Scales), developed by Antony et al. (1998), is the most frequently utilized psychological scale for depression ($k = 8$) and anxiety ($k = 8$). The Satisfaction with Life Scale (SWLS) for quality of life ($k = 13$), Buss-Perry Aggression Questionnaire (BPAQ) for aggression ($k = 8$), UCLA Loneliness Scale for loneliness ($k = 16$) and Young's Internet Addiction Test ($k = 10$) for internet addiction were also frequently utilized (Russell et al., 1980; Diener et al., 1985; Buss and Perry, 1992; Young, 1998).

The results of the quantitative synthesis for the association between GD and other variables are shown in **Table 3**. The overall estimated mean effect sizes of the psychological variables for GD are as follows: Depression ($r = 0.33$), anxiety ($r = 0.29$), aggression ($r = 0.30$), QOL ($r = -0.22$), and loneliness ($r = 0.29$). The estimated effect sizes of internet addiction and game playtime are $r = 0.56$ and $r = 0.40$. The forest plots displaying the population estimate and the effect sizes of individual studies for each of the variables are presented in **Figures 2**–**4**.

Heterogeneity and Moderator Analyses

The results of the quantitative synthesis indicated high levels of heterogeneity for all the variables. The heterogeneity estimates are presented in **Table 3**. Categorical moderator analyses and meta regression analyses using moderators were conducted to identify the potential sources of heterogeneity. Moderator (a), the specific GD instrument used in the study, (categorized into "IGDS9-SF," "GAS-7," "Lemmens IGD-9," "AICA," and "IGDT-10"), moderator (b), the type of the sample (categorized into "adolescents," "adults," and "both"), and (c) the study location (categorized into six continents) were used as the moderators if each of the subgroups had sufficient number of studies (Fu et al., 2011). Moderator (a) was a significant moderator for anxiety and GD ($p = 0.02$), and moderator (c) was a significant moderator for aggression and GD ($p < 0.01$). Moderator (d), gender ratio of the participants of each study, was a significant moderator only for game playtime ($p = 0.04$), indicating that the studies having more male participants reported smaller correlation coefficients between GD and game playtime. The results of the categorical and continuous moderator analysis of validity generalization are presented in the (**Supplementary Tables 2**, **3**).

Publication Bias

Publication bias for validity generalization was assessed by using funnel plots and Egger's regression test. The funnel plots

TABLE 3 | Association between GD and psychological/behavioral variables.

					95% CI	Heterogeneity		
Psychological variables	**j**	**k**	**n**	r_{obs}	**LL, UL**	**τ**	**τ^2**	**I^2(%)**
Depression	44	45	83,604	0.33	[0.29, 0.36]	0.14	0.019	95.6
Anxiety	36	37	76,948	0.29	[0.25, 0.33]	0.13	0.016	97.2
Aggression	17	19	35,441	0.30	[0.24, 0.35]	0.13	0.017	96.9
QOL	17	18	25,833	–0.22	[–0.31, –0.12]	0.21	0.043	96.1
Loneliness	17	18	26,677	0.29	[0.22, 0.36]	0.16	0.027	95.8
Internet addiction	18	20	25,368	0.56	[0.48, 0.63]	0.25	0.062	98.2
Game playtime	51	53	62,792	0.40	[0.35, 0.45]	0.22	0.048	97.6

j, number of studies; k, number of reported effect sizes; n, number of samples; r_{obs}, estimated mean effect sizes (correlation coefficient); SD_r, standard deviation for r_{obs}; CI, confidence interval; τ, square root of estimated tau^2; τ^2, estimated amount of total heterogeneity; I^2, total heterogeneity/total variability.

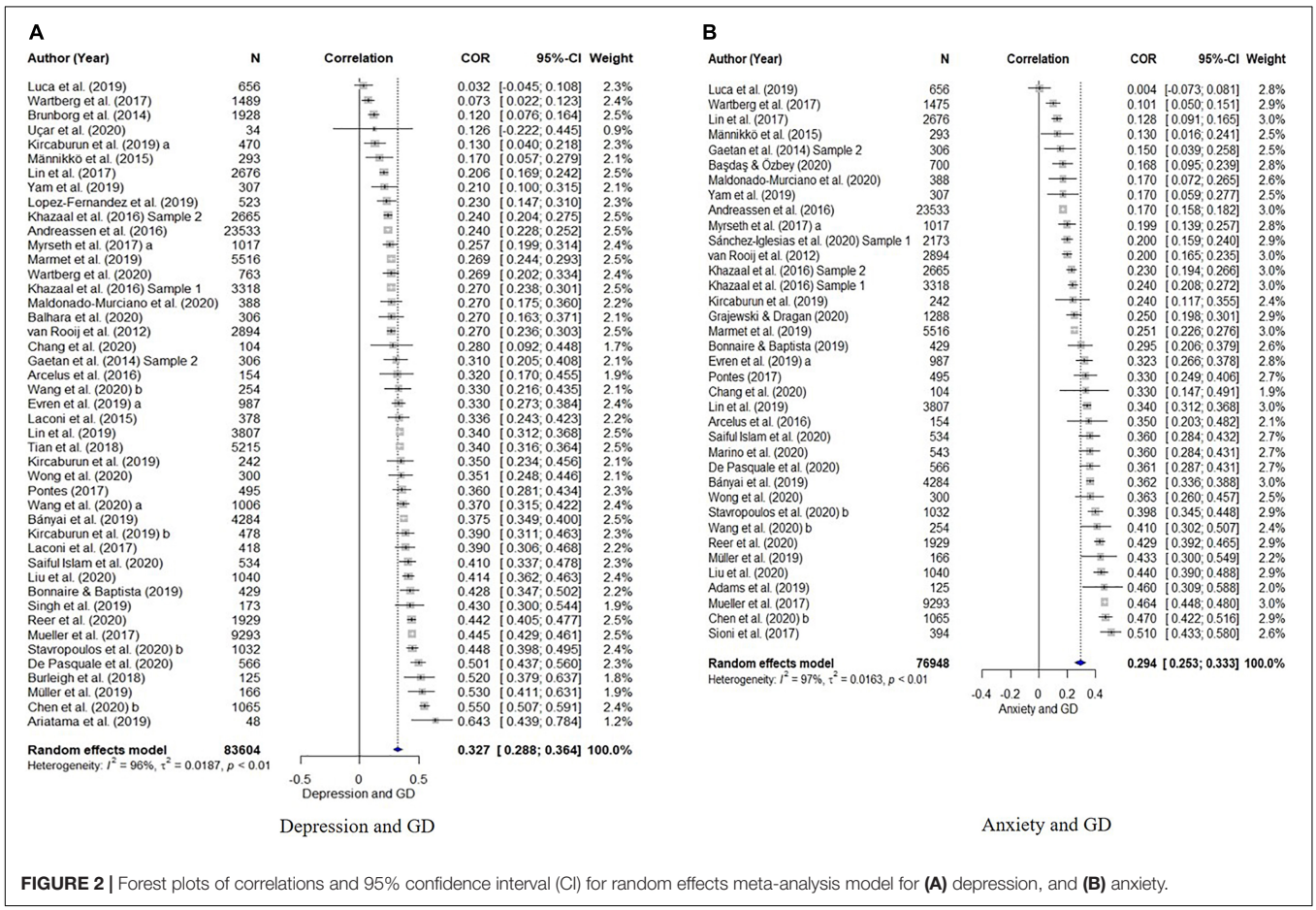

FIGURE 2 | Forest plots of correlations and 95% confidence interval (CI) for random effects meta-analysis model for **(A)** depression, and **(B)** anxiety.

for all the variables have been visualized in **Supplementary Figure 3**. Since visual inspection can be subjective, Egger's regression tests for the detection of funnel plot asymmetry were performed (Sterne et al., 2000). The results of the regression tests for game play time were statistically significant (t = 3.16, p < 0.01), suggesting the presence of evidence for publication bias. Cumulative meta-analysis and sensitivity analysis were further conducted to investigate the publication bias of studies reporting the correlation between GD and game playtime. The results of the cumulative meta-analysis and sensitivity analysis revealed that the studies conducted by, and Brunborg et al. (2014) and Bányai et al. (2019) had influenced the overall effect size estimate as two studies reported exceptionally small and large effect sizes. Omitting study by Brunborg et al. (2014) decreased the overall effect size estimate between GD and game playtime to r = 0.39 while omitting study by Bányai et al. (2019) increased the overall effect size estimate to r = 0.41. The result of sensitivity analysis for GD and game playtime is provided in **Figure 5**. The results of Egger's regression test for the other variables were insignificant (for depression t = 0.98, p = 0.33; for anxiety t = 1.02,

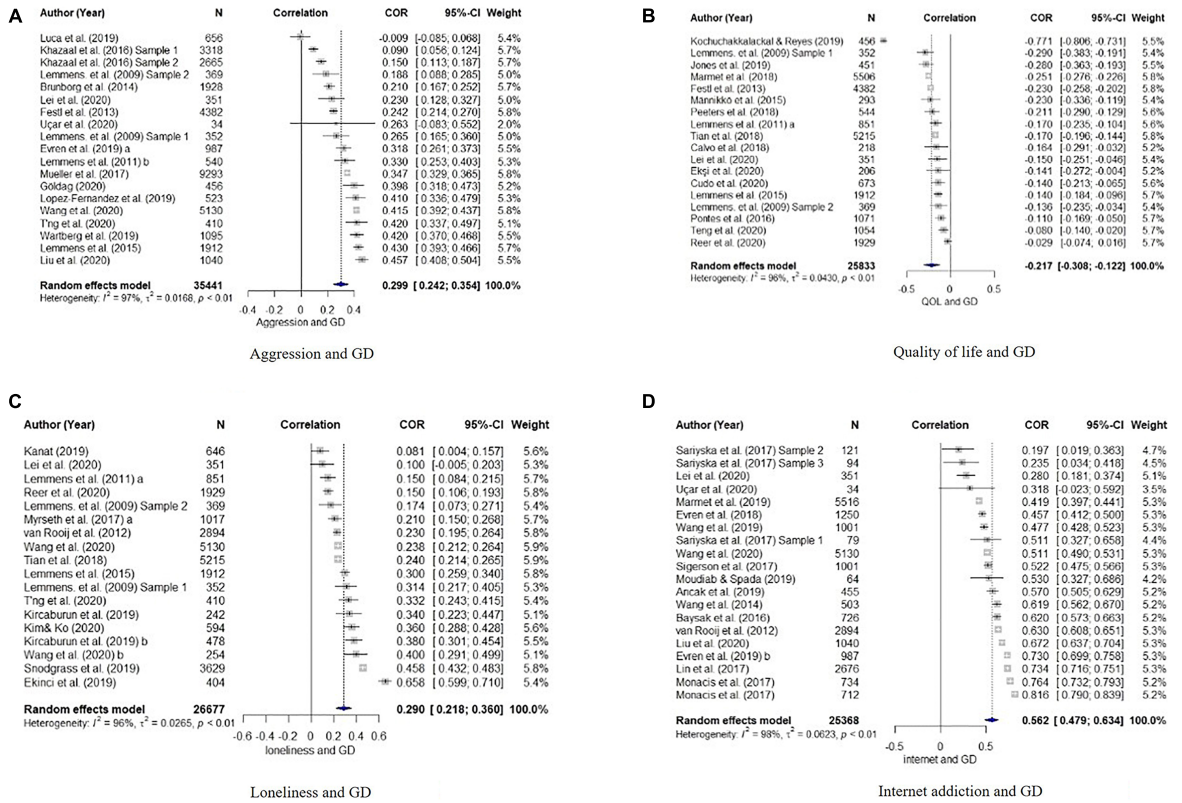

FIGURE 3 | Forest plots of correlations and 95% confidence interval (CI) for random effects meta-analysis model for **(A)** aggression, **(B)** quality of life, **(C)** loneliness, and **(D)** internet addiction.

p = 0.31; for aggression t = –0.23, p = 0.82; for QOL t = –0.37, p = 0.72; for loneliness t = 0.33, p = 0.75; for internet addiction t = 0.49, p = 0.63).

DISCUSSION

Reliability

The current study aimed to provide information on what GD scales measure, and how consistent the measure is. The current study conducted meta-analyses by quantitatively synthesizing the Cronbach's alpha reliability coefficients and bivariate Pearson's correlations. The result of the quantitative synthesis of alpha coefficients, reliability generalization, showed an estimated alpha coefficient of 0.86. A high value of alpha coefficient is usually desirable (Cronbach, 1951), but an alpha coefficient above 0.9 may indicate unnecessary redundancy rather than a desirable level of internal consistency (Streiner, 2003). In this regard, the estimated alpha coefficient of 0.86 can be interpreted as an indication of good internal consistency (Gliem and Gliem, 2003). With respect to the moderator analysis, each tool displayed Cronbach's alpha coefficients ranged from 0.81 to 0.89. The 172 studies in total presented 193 effect sizes of alpha coefficients as the measures of internal consistency. Alpha coefficients of studies with IGDS9-SF were most frequently reported, and the result of ANOVA revealed that IGDS9-SF possesses the highest estimated alpha followed by AICA. The funnel plot and Egger's test of each GD tool indicated the existence of a potential publication bias for GAS-7 (z = –2.02, p = 0.04). The funnel plots for the GD tools are provided in the (**Supplementary Figure 2**).

Given that the current study only included the psychometrically sound tools to synthesize the reliability coefficients, there is a possibility that the reliability estimation of the current study might be positively biased. A categorical moderator analysis with the specific GD instrument used in the study, was performed to examine whether there were differences between each GD tool. The results of the omnibus subgroup test rejected the null hypothesis, indicating that there are differences between the estimated alpha coefficients of each of the tools. ANOVA analyses between every two GD tools were further performed as the omnibus test results increase the type 1 error. The results indicated that IGDS9-SF (α = 0.89) had the highest estimated alpha, followed by AICA (α = 0.85). Lemmens IGD-9 showed the lowest estimated alpha (α = 0.81) among all the tools.

Caution should be taken in interpreting the results of the pooled Cronbach's alpha coefficient. The high Cronbach's alpha is not a perfect index of internal consistency as alpha by itself does not assure an excellent degree of internal consistency (Tavakol and Dennick, 2011). An alpha coefficient can be susceptible to the length of the test, undue narrowness

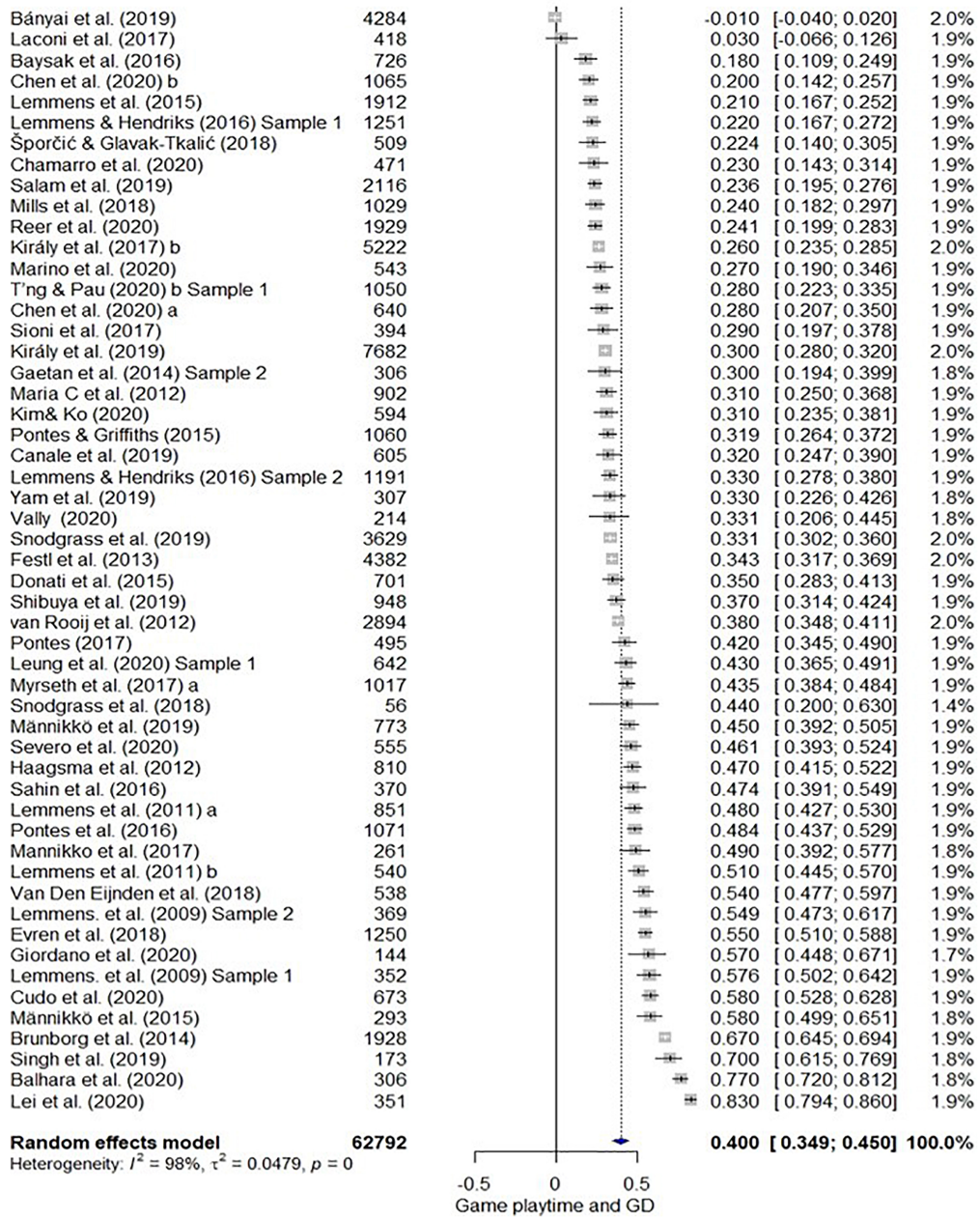

FIGURE 4 | Forest plot of correlations and 95% confidence interval (CI) for random effects meta- analysis model for game playtime.

(Streiner, 2003), and dimensionality (Green et al., 1977). The test-retest reliability coefficients can provide additional information on overall reliability when they are interpreted together with the internal consistency coefficients. An intraclass correlation coefficient or test-retest interval correlation coefficient can be referred as the stability or reproducibility of the test (Polit, 2014). The estimated average of the eight reliability coefficients was 0.86 (95% CI = 0.81–0.89) which can be interpreted as a good level (Cicchetti, 1994). More studies should examine the test-retest reliability of GD assessment tools as a very small number of studies have reported on retest reliability in comparison to the studies that have reported on internal consistency.

Validity and Association

The bivariate Pearson's correlation between the seven variables and GD tools were coded. The estimated effect sizes of the correlation ranged between 0.22 and 0.56 in magnitude. The estimated associations between GD and psychological/behavioral variables were found. The Hedge's

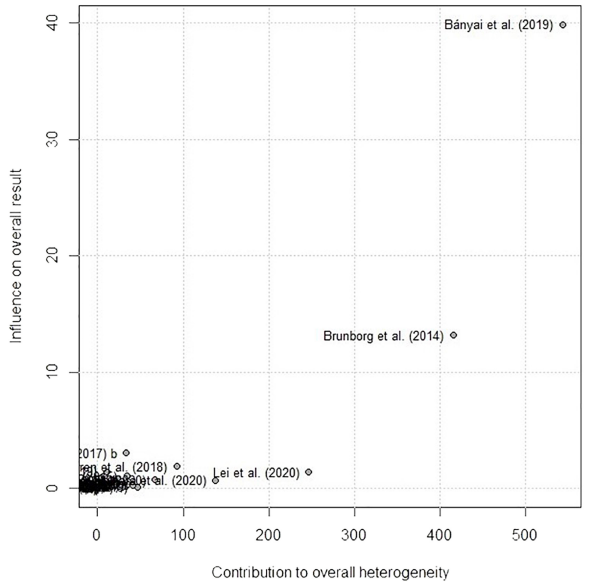

FIGURE 5 | Sensitivity analysis for game playtime and GD.

estimator (Borenstein et al., 2011) for the seven variables are as follows: 0.33 for depression, 0.29 for anxiety, 0.30 for aggression, –0.22 for QOL, 0.29 for loneliness, 0.56 for internet addiction, and 0.40 for game playtime.

By synthesizing the effect size of correlation coefficients and examining the convergent and discriminant validity of GD tools, we aimed to scrutinize the association between GD and mental disorders. Unfortunately, the current study offers information only on the association, rather than on causality. The results from the current study do not suggest that the correlation effect sizes are small or large enough to help the society make clear distinction. Since the labeling of the effect size magnitude can be arbitrary (Schober et al., 2018), we suggest an interpretation of the results by comparing each of the effect sizes. For instance, GD tools have a correlation effect size of 0.40 with game playtime and 0.33 with depression, meaning that the depression was found to have a slightly smaller association with GD than the gaming behavior. Anxiety (r = 0.29), aggression (r = 0.30), and loneliness (r = 0.29) showed similar magnitudes of correlation effect sizes. QOL was the only variable negatively associated with GD (r = –0.22). Internet addiction showed the highest correlation with GD. The overlapped items between internet addiction and gaming disorder, especially the IGD criteria for DSM-5, might contribute toward a high association between internet addiction and GD.

The results of the moderator analysis show that the specific GD instrument used in the study significantly moderates the correlation between anxiety and GD. IGDS9-SF captures higher associations (r = 0.33) between anxiety and GD than GAS-7 (r = 0.23). This might be due to the different features of each scales. Study location was found to be a significant moderator for the correlation between aggression and GD. The studies conducted in Asia reported higher association (r = 0.38) between aggression and GD than the studies conducted in Europe (r = 0.24). This is consistent with the findings of previous studies. Studies reporting the role of aggression in gaming disorders have investigated the mediating role of ethnicity and cultural differences (Kim et al., 2018; Prescott et al., 2018). Anderson et al. (2010) also reported that cultural difference can moderate the association between violence, prosocial behavior, and video gaming. A continuous variable moderator analysis shows that the gender ratio of study participants was a significant moderating continuous variable. The higher the percentage of female participants, the stronger the association between game playtime and GD (b = 0.6302 for intercept; b = –0.0033 for one percent point increase in the percent of male participants). The males are known to be more vulnerable than females in developing a gaming disorder (Dong et al., 2018; Fam, 2018). The game playtime seems to have a more direct effect on females than on males.

The Egger's test, cumulative meta-analysis, and sensitivity analysis revealed an asymmetry in the publications reporting the correlations between game playtime and GD. The studies conducted by Brunborg et al. (2014) and Bányai et al. (2019) influenced the overall effect size. Notably, Bányai et al. (2019) reported Pearson's bivariate correlation between game playtime and GD of r = –0.01, which is in essence zero. Since the study by Bányai et al. (2019) included e-sport gamers who spent significantly more time playing games than recreational gamers, the correlation reported by the author significantly differs from that of the other studies. The findings of Bányai et al. (2019) presented the moderating role of gaming motivation in causing GD and psychiatric distress, indicating that gaming behavior itself can have even no association with the GD.

The main findings of the current study show that the magnitudes of the effect sizes of convergent and discriminant validities of GD are not significantly different. Given the association of 0.40 between game playtime and GD, common symptoms (e.g., depression, anxiety) of psychopathology also showed considerable associations with GD. As González-Bueso et al. (2018) commented, we agree to the idea that whether the problematic gaming behaviors are a consequence, or a trigger of other psychopathologies cannot be unraveled yet. Studies have reported that just as problematic gaming increases psychological distress, psychological factors such as low self-esteem and loneliness also bidirectionally affect or predict problematic gaming (Lemmens et al., 2011; Tian et al., 2017; Tras, 2019; Wartberg et al., 2019).

To identify the unraveled relationship between GD and psychopathology, and move beyond these debates, future studies must come to a consensus on the diagnostic criteria of gaming disorder. Delphi method can be helpful in developing the diagnostic criteria of GD and arriving at a consensus (Castro-Calvo et al., 2021). The tools should be improved and unified rather than continuously developed by various researchers. Importantly, the clinician interview must be adopted in this field to verify the positive cases of GD and report comorbid psychopathologies (Pontes and Griffiths, 2019). Of the 184 studies included in the current meta-analysis study, only nine studies included clinical samples and adopted structured clinician interviews in a strict sense (e.g., Müller et al., 2019;

Wölfling et al., 2019; Phan et al., 2020). Longitudinal and high-quality clinical trial studies (e.g., Han et al., 2017; Li et al., 2017; Wölfling et al., 2019) are also necessary to rebut the argument that the problematic gaming behavior is a consequence of other psychopathologies. With respect to the other aspects of validity, future studies should actively examine the predictive validity using gold standard tool of the diagnosis.

Study Limitations

Some limitations should be noted. First, despite our effort to include all the relevant studies, some could not be coded owing to unreported data. To minimize this limitation, we reached out to researchers, and received relevant information from 17 researchers. Second, the current study focuses on the five GD assessment tools recommend by King et al. (2020). Since more than 40 assessment tools have been developed to assess GD, the representativeness of the five tools included in the current study could be questioned. Rather than establishing our own selection criteria, we selected the five GD assessment tools based on a rigorous review article by King et al. (2020). The third limitation might reside in the conventional two-level meta-analysis model and the high level of study heterogeneity found in both reliability and validity generalization. While efforts were made to investigate the potential reason for high heterogeneity, the categorical and continuous moderator analysis only partially adjusted the heterogeneity. We adopted the conventional two-level meta-analysis model instead of three-level model or robust variance estimation method due to scarce report of the variance of the individual effect sizes within each study. We used effect sizes from longitudinal studies (k = 17) and several effect sizes reported from the same sample (k = 3), and those effect-sizes reported from the same study were not analyzed repeatedly in the current study. If variance of the individual effect sizes within each study are accumulated in a future, a three-level meta-analysis model or robust estimation technique would be recommended to handle the dependent effect sizes and considering within- and between-study heterogeneity. The fourth limitation is that due to insufficient number of studies, we did not perform a meta-analysis for GD and attention deficit hyperactivity disorder, which is a common psychiatric comorbidity in clinical practice (Yen et al., 2017). Five studies reported Pearson's correlation coefficients ranging from 0.16 to 0.38 between GD and impulsivity. Given the high heterogeneity, we decided that the number of studies on impulsivity was insufficient to carry out a meta-analysis. Fifth, since majority of the included studies in the current study adopted either GAS-7 and IGDS9-SF, the feature of the GAS-7 and IGDS9-SF might affect the effect size estimation. The limitation should be addressed as more studies in this field are conducted.

CONCLUSION

Despite its limitations, this is the first and largest systematic review study (with 184 studies and 285,752 study participants) to examine the association between GD and psychological/behavioral variables by synthesizing the reliability, and convergent and discriminant validity information of the five GD assessment tools (e.g., IGDS9-SF, GAS-7, Lemmens IGD-9, AICA, and IGDT-10). In addition to the reliability generalization of the GD assessment tools, a major strength of this study is that we applied meta-analytic techniques to investigate the magnitude of relationships between GD and common symptoms of mental disorders (e.g., depression, anxiety disorders, addictions, impulsivity, and hostility), as indicated in previous studies (Han et al., 2017; Na et al., 2017; Wang et al., 2017; González-Bueso et al., 2018; Liu et al., 2018). We also applied same meta-analytic technique to examine the magnitude of association between GD and the gaming behavior. We believe that this meta-analysis provides current status of GD. Future studies should address debatable issues in reliability and convergent/discriminant validity of the GD assessment tools, and more studies should be conducted to better understand the bidirectional relationship between GD and other psychopathologies.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

SY, W-YA, JK, S-HS, JC, and K-HC contributed to the conception and design of the study. SY, YY, and ER coded the data and wrote the first draft of the manuscript. SY, YY, ER, and K-HC double-checked the coded data. SY and YY analyzed data. K-HC supervised the overall study process. W-YA, JK, S-HS, JC, and K-HC contributed editing the draft of the manuscript. All authors have read and approved the submitted manuscript.

FUNDING

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C2099665). The study was also supported by the Investigating Scientific Evidence for Registering Gaming Disorder on Korean Standard Classification of Disease and Cause of Death project, of Ministry of Health and Welfare, Korea, and Korea Creative Content Agency.

ACKNOWLEDGMENTS

We would like to thank Editage (www.editage.co.kr) for English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.764209/full#supplementary-material

REFERENCES

Aarseth, E., Bean, A. M., Boonen, H., Colder Carras, M., Coulson, M., Das, D., et al. (2017). Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal. *J. Behav. Addict.* 6, 267–270. doi: 10.1556/2006.5.2016.088

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders*, 5th Edn. Arlington, VA: American Psychiatric Publishing. doi: 10.1176/appi.books.9780890425596

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., et al. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: a meta-analytic review. *Psychol. Bull.* 136, 151–173. doi: 10.1037/a0018251

Andreetta, J., Teh, J., Burleigh, T. L., Gomez, R., and Stavropoulos, V. (2020). Associations between comorbid stress and internet gaming disorder symptoms: are there cultural and gender variations? *Asia Pac. Psychiatry* 12:e12387. doi: 10.1111/appy.12387

Antony, M. M., Bieling, P. J., Cox, B. J., Enns, M. W., and Swinson, R. P. (1998). Psychometric properties of the 42-item and 21-item versions of the depression anxiety stress scales in clinical groups and a community sample. *Psychol. Assess.* 10, 176–181. doi: 10.1037/1040-3590.10.2.176

Assink, M., and Wibbelink, C. J. (2016). Fitting three-level meta-analytic models in R: A step-by-step tutorial. *Quant. Methods Psychol.* 12, 154–174. doi: 10.20982/tqmp.12.3.p154

Bányai, F., Griffiths, M. D., Demetrovics, Z., and Király, O. (2019). The mediating effect of motivations between psychiatric distress and gaming disorder among esport gamers and recreational gamers. *Compr. Psychiatry* 94:152117. doi: 10.1016/j.comppsych.2019.152117

Billieux, J., King, D. L., Higuchi, S., Achab, S., Bowden-Jones, H., Hao, W., et al. (2017). Functional impairment matters in the screening and diagnosis of gaming disorder: commentary on: Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal (Aarseth et al.). *J. Behav. Addict.* 6, 285–289.

Billieux, J., Schimmenti, A., Khazaal, Y., Maurage, P., and Heeren, A. (2015). Are we overpathologizing everyday life? A tenable blueprint for behavioral addiction research. *J. Behav. Addict.* 4, 119–123. doi: 10.1556/2006.4.2015.009

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *J. Educ. Behav. Stat.* 27, 335–340. doi: 10.3102/10769986027004335

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods.* 1, 97–111. doi: 10.1002/jrsm.12

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2011). *Introduction to Meta-Analysis*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470743386

Borenstein, M., Higgins, J. P., Hedges, L. V., and Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Res. Synth. Methods.* 8, 5–18. doi: 10.1002/jrsm.1230

Brannick, M. T., Potter, S. M., Benitez, B., and Morris, S. B. (2019). Bias and precision of alternate estimators in meta-analysis: benefits of blending schmidt-hunter and hedges approaches. *Organ. Res. Methods.* 22, 490–514. doi: 10.1177/1094428117741966

Brunborg, G. S., Mentzoni, R. A., and Frøyland, L. R. (2014). Is video gaming, or video game addiction, associated with depression, academic achievement, heavy episodic drinking, or conduct problems? *J. Behav. Addict.* 3, 27–32. doi: 10.1556/jba.3.2014.002

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

Castro-Calvo, J., King, D. L., Stein, D. J., Brand, M., Carmi, L., Chamberlain, S. R., et al. (2021). Expert appraisal of criteria for assessing gaming disorder: an international delphi study. *Addiction* 116, 2463–2475. doi: 10.1111/add.15411

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychol. Assess.* 6, 284–290. doi: 10.1037/1040-3590.6.4.284

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*. 16, 297–334. doi: 10.1007/BF02310555

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

Dong, G., Wang, L., Du, X., and Potenza, M. N. (2018). Gender-related differences in neural responses to gaming cues before and after gaming: implications for gender-specific vulnerabilities to Internet gaming disorder. *Soc. Cogn. Affect. Neurosci.* 13, 1203–1214. doi: 10.1093/scan/nsy084

Egger, M., Smith, G. D., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj* 315, 629–634. doi: 10.1136/bmj.315.7109.629

Fam, J. Y. (2018). Prevalence of internet gaming disorder in adolescents: a meta-analysis across three decades. *Scand. J. Psychol.* 59, 524–531. doi: 10.1111/sjop.12459

Fauth-Bühler, M., and Mann, K. (2017). Neurobiological correlates of internet gaming disorder: similarities to pathological gambling. *Addict. Behav.* 64, 349–356. doi: 10.1016/j.addbeh.2015.11.004

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychol. Methods*. 10, 444–467. doi: 10.1037/1082-989X.10.4.444

Field, A. P., and Gillett, R. (2010). How to do a meta-analysis. *Br. J. Math. Stat. Psychol.* 63, 665–694. doi: 10.1348/000711010X502733

Fu, R., Gartlehner, G., Grant, M., Shamliyan, T., Sedrakyan, A., Wilt, T. J., et al. (2011). Conducting quantitative synthesis when comparing medical interventions: AHRQ and the effective health care program. *J. Clin. Epidemiol.* 64, 1187–1197. doi: 10.1016/j.jclinepi.2010.08.0100

Gliem, J. A., and Gliem, R. R. (2003). "Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales," in *Proceedings of the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education*, Columbus, OH.

González-Bueso, V., Santamaría, J. J., Fernández, D., Merino, L., Montero, E., and Ribas, J. (2018). Association between internet gaming disorder or pathological video-game use and comorbid psychopathology: a comprehensive review. *Int. J. Environ. Res. Public Health* 15:668. doi: 10.3390/ijerph15040668

Green, S. B., Lissitz, R. W., and Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality1. *Educ. Psychol. Meas.* 37, 827–838. doi: 10.1177/001316447703700403

Griffiths, M. D., Kuss, D. J., Lopez-Fernandez, O., and Pontes, H. M. (2017). Problematic gaming exists and is an example of disordered gaming: commentary on: Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal (Aarseth et al.). *J. Behav. Addict.* 6, 296–301. doi: 10.1556/2006.6.2017.037

Han, D. H., Kim, S. M., Bae, S., Renshaw, P. F., and Anderson, J. S. (2017). Brain connectivity and psychiatric comorbidity in adolescents with internet gaming disorder. *Addict. Biol.* 22, 802–812. doi: 10.1111/adb.12347

Han, D. H., Yoo, M., Renshaw, P. F., and Petry, N. M. (2018). A cohort study of patients seeking internet gaming disorder treatment. *J. Behav. Addict.* 7, 930–938. doi: 10.1556/2006.7.2018.102

Harrer, M., Cuijpers, P., Furukawa, T. A., and Ebert, D. D. (2021). *Doing Meta-Analysis With R: A Hands-On Guide*. Boca Raton, FL: Chapmann & Hall/CRC Press.

Harrer, M., Cuijpers, P., Furukawa, T., and Ebert, D. (2019). *Doing Meta-Analysis in R: A Hands-on Guide*. Available online at: https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R (accessed August 25, 2021).

Hedges, L. V. (1992). Meta-analysis. *J. Educ. Stat.* 17, 279–296. doi: 10.3102/10769986017004279

Hedges, L. V., Tipton, E., and Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Res. Synth. Methods* 1, 39–65. doi: 10.1002/jrsm.5

Hunter, J. E., and Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage Publishing.

Jeong, H., Yim, H. W., Lee, S.-Y., Lee, H. K., Potenza, M. N., Kwon, J.-H., et al. (2018). Discordance between self-report and clinical diagnosis of internet gaming disorder in adolescents. *Sci. Rep.* 8, 1–8. doi: 10.1038/s41598-018-28478-8

Jo, Y. S., Bhang, S. Y., Choi, J. S., Lee, H. K., Lee, S. Y., and Kweon, Y.-S. (2019). Clinical characteristics of diagnosis for internet gaming disorder: comparison of DSM-5 IGD and ICD-11 GD diagnosis. *J. Clin. Med.* 8, 945–957. doi: 10.3390/jcm8070945

Jones, C., Scholes, L., Johnson, D., Katsikitis, M., and Carras, M. C. (2014). Gaming well: links between videogames and flourishing mental health. *Front. Psychol.* 5:260–267. doi: 10.3389/fpsyg.2014.00260

Kardefelt-Winther, D. (2014). Problematizing excessive online gaming and its psychological predictors. *Comput. Hum. Behav.* 31, 118–122. doi: 10.1016/j.chb.2013.10.017

Kim, E., Yim, H. W., Jeong, H., Jo, S.-J., Lee, H. K., Son, H. J., et al. (2018). The association between aggression and risk of Internet gaming disorder in Korean adolescents: the mediation effect of father-adolescent communication style. *Epidemiol. Health* 40:e2018039. doi: 10.4178/epih.e2018039

King, D. L., Chamberlain, S. R., Carragher, N., Billieux, J., Stein, D., Mueller, K., et al. (2020). Screening and assessment tools for gaming disorder: a comprehensive systematic review. *Clin. Psychol. Rev.* 77:101831. doi: 10.1016/j.cpr.2020.101831

King, D. L., Delfabbro, P. H., Wu, A. M., Doh, Y. Y., Kuss, D. J., Pallesen, S., et al. (2017). Treatment of internet gaming disorder: an international systematic review and CONSORT evaluation. *Clin. Psychol. Rev.* 54, 123–133. doi: 10.1016/j.cpr.2017.04.002

Király, O., and Demetrovics, Z. (2017). Inclusion of gaming disorder in ICD has more advantages than disadvantages: commentary on: scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal (Aarseth et al.). *J. Behav. Addict.* 6, 280–284. doi: 10.1556/2006.6.2017.046

Király, O., Sleczka, P., Pontes, H. M., Urbán, R., Griffiths, M. D., and Demetrovics, Z. (2017). Validation of the ten-item Internet Gaming Disorder Test (IGDT-10) and evaluation of the nine DSM-5 Internet Gaming Disorder criteria. *Addict. Behav.* 64, 253–260. doi: 10.1016/j.addbeh.2015.11.005

Kräplin, A., Scherbaum, S., Kraft, E.-M., Rehbein, F., Bühringer, G., Goschke, T., et al. (2021). The role of inhibitory control and decision-making in the course of Internet Gaming Disorder. *J. Behav. Addict.* 9, 990–1001. doi: 10.1556/2006.2020.00076

Kuss, D. J., Griffiths, M. D., and Pontes, H. M. (2017). DSM-5 diagnosis of Internet Gaming Disorder: Some ways forward in overcoming issues and concerns in the gaming studies field: Response to the commentaries. *J. Behav. Addict.* 6, 133–141. doi: 10.1556/2006.6.2017.032

Lemmens, J. S., Valkenburg, P. M., and Gentile, D. A. (2015). The internet gaming disorder scale. *Psychol. Assess.* 27:567. doi: 10.1037/pas0000062

Lemmens, J. S., Valkenburg, P. M., and Peter, J. (2009). Development and validation of a game addiction scale for adolescents. *Media Psychol.* 12, 77–95. doi: 10.1080/15213260802669458

Lemmens, J. S., Valkenburg, P. M., and Peter, J. (2011). Psychosocial causes and consequences of pathological gaming. *Comput. Hum. Behav.* 27, 144–152. doi: 10.1016/j.chb.2010.07.015

Li, W., Garland, E. L., McGovern, P., O'Brien, J. E., Tronnier, C., and Howard, M. O. (2017). Mindfulness-oriented recovery enhancement for internet gaming disorder in US adults: a stage I randomized controlled trial. *Addict Behav.* 31, 393–402. doi: 10.1037/adb0000269

Liu, L., Yao, Y.-W., Li, C.-s. R., Zhang, J.-T., Xia, C.-C., Lan, J., et al. (2018). The comorbidity between internet gaming disorder and depression: interrelationship and neural mechanisms. *Front. Psychiatry* 9:154. doi: 10.3389/fpsyt.2018.00154

López-Pina, J. A., Sánchez-Meca, J., López-López, J. A., Marín-Martínez, F., Núñez-Núñez, R. M., Rosa-Alcázar, A. I., et al. (2015). The Yale–Brown obsessive compulsive scale: a reliability generalization meta-analysis. *Assessment* 22, 619–628. doi: 10.1177/1073191114551954

Miller, B. K., Nicols, K. M., Clark, S., Daniels, A., and Grant, W. (2018). Meta-analysis of coefficient alpha for scores on the narcissistic personality inventory. *PLoS One* 13:e0208331. doi: 10.1371/journal.pone.0208331

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 6:e1000097. doi: 10.1371/journal.pmed.1000097

Müller, K. W., Beutel, M. E., and Wölfling, K. (2014). A contribution to the clinical characterization of Internet addiction in a sample of treatment seekers: validity of assessment, severity of psychopathology and type of co-morbidity. *Compr. Psychiatry* 55, 770–777. doi: 10.1016/j.comppsych.2014.01.010

Müller, K. W., Beutel, M. E., Dreier, M., and Wölfling, K. (2019). A clinical evaluation of the DSM-5 criteria for internet gaming disorder and a pilot study on their applicability to further Internet-related disorders. *J. Behav. Addict.* 8, 16–24. doi: 10.1556/2006.7.2018.140

Na, E., Lee, H., Choi, I., and Kim, D. J. (2017). Comorbidity of Internet Gaming Disorder and alcohol use disorder: a focus on clinical characteristics and gaming patterns. *Am. J. Addict.* 26, 326–334. doi: 10.1111/ajad.12528

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj* 372:71. doi: 10.1136/bmj.n71

Petry, N. M., and O'Brien, C. P. (2013). Internet gaming disorder and the DSM-5. *Addiction* 108, 1186–1187. doi: 10.1111/add.12162

Phan, O., Prieur, C., Bonnaire, C., and Obradovic, I. (2020). Internet gaming disorder: exploring its impact on satisfaction in life in PELLEAS adolescent Sample. *Int. J. Environ. Res. Public Health*. 17:3. doi: 10.3390/ijerph17010003

Polit, D. F. (2014). Getting serious about test–retest reliability: a critique of retest research and some recommendations. *Qual. Life Res.* 23, 1713–1720. doi: 10.1007/s11136-014-0632-9

Pontes, H. M., and Griffiths, M. D. (2015). Measuring DSM-5 internet gaming disorder: development and validation of a short psychometric scale. *Comput. Hum. Behav.* 45, 137–143. doi: 10.1016/j.chb.2014.12.006

Pontes, H. M., and Griffiths, M. D. (2019). A new era for gaming disorder research: Time to shift from consensus to consistency. *Addict. Behav.* 103:106059. doi: 10.1016/j.addbeh.2019.106059

Pontes, H. M., Macur, M., and Griffiths, M. D. (2016). Internet gaming disorder among slovenian primary schoolchildren: findings from a nationally representative sample of adolescents. *J. Behav. Addict.* 5, 304–310. doi: 10.1556/2006.5.2016.042

Prescott, A. T., Sargent, J. D., and Hull, J. G. (2018). Metaanalysis of the relationship between violent video game play and physical aggression over time. *Proc. Natl. Acad. Sci. U.S.A.* 115, 9882–9888. doi: 10.1073/pnas.1611617114

Quintana, D. S. (2015). From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. *Front. Psychol.* 6:1549. doi: 10.3389/fpsyg.2015.01549

Rodriguez, M. C., and Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychol. Methods.* 11, 306–322. doi: 10.1037/1082-989X.11.3.306

Rönkkö, M., and Cho, E. (2020). An updated guideline for assessing discriminant validity. *Organ. Res. Methods* 2:1094428120968614. doi: 10.1177/1094428120968614

Rothstein, H. R., Sutton, A. J., and Borenstein, M. (2005). *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0470870168

Rumpf, H.-J., Achab, S., Billieux, J., Bowden-Jones, H., Carragher, N., Demetrovics, Z., et al. (2018). Including gaming disorder in the ICD-11: The need to do so from a clinical and public health perspective: Commentary on: a weak scientific basis for gaming disorder: Let us err on the side of caution (van Rooij et al., 2018). *J. Behav. Addict.* 7, 556–561. doi: 10.1556/2006.7.2018.59

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

Sariyska, R., Lachmann, B., Markett, S., Reuter, M., and Montag, C. (2017). Individual differences in implicit learning abilities and impulsive behavior in the context of internet addiction and Internet Gaming Disorder under the consideration of gender. *Addict. Behav. Rep* 5, 19–28. doi: 10.1016/j.abrep.2017.02.002

Schmidt, F. L., and Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychol. Bull.* 124:262. doi: 10.1037/0033-2909.124.2.262

Schober, P., Boer, C., and Schwarte, L. A. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* 126, 1763–1768. doi: 10.1213/ANE.0000000000002864

Schwarzer, G. (2007). Meta: an R package for meta-analysis. *R News* 7, 40–45. doi: 10.1007/978-3-319-21416-0

Sterne, J. A., Egger, M., and Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *Bmj* 323, 101–105. doi: 10.1136/bmj.323.7304.101

Sterne, J. A., Gavaghan, D., and Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *J. Clin. Epidemiol.* 53, 1119–1129. doi: 10.1016/S0895-4356(00)00242-0

Stevens, M. W., Dorstyn, D., Delfabbro, P. H., and King, D. L. (2021). Global prevalence of gaming disorder: a systematic review and meta-analysis. *Aust. N. Z. J. Psychiatry* 55, 553–568. doi: 10.1177/0004867420962851

Streiner, D. L. (2003). Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J. Pers. Assess.* 80, 99–103. doi: 10.1207/S15327752JPA8001_18

Tavakol, M., and Dennick, R. (2011). Making sense of Cronbach's alpha. *Int. J. Med. Educ.* 2, 53–55. doi: 10.5116/ijme.4dfb.8dfd

Tian, Y., Bian, Y., Han, P., Gao, F., and Wang, P. (2017). Associations between psychosocial factors and generalized pathological internet use in Chinese university students: a longitudinal cross-lagged analysis. *Comput. Hum. Behav.* 72, 178–188. doi: 10.1016/j.chb.2017.02.048

Tras, Z. (2019). Internet addiction and loneliness as predictors of Internet Gaming Disorder in adolescents. *Educ. Res. Rev.* 14, 465–473. doi: 10.5897/ERR2019.3768

Vacha-Haase, T. (1998). Reliability generalization: exploring variance in measurement error affecting score reliability across studies. *Educ. Psychol. Meas.* 58, 6–20. doi: 10.1177/0013164498058001002

Van Den Brink, W. (2017). ICD-11 gaming disorder: needed and just in time or dangerous and much too early? Commentary on: Scholars' open debate paper on the World Health Organization ICD-11 Gaming Disorder proposal (Aarseth et al.). *J. Behav. Addict.* 6, 290–292. doi: 10.1556/2006.6.2017.040

Van Rooij, A. J., Ferguson, C. J., Colder Carras, M., Kardefelt-Winther, D., Shi, J., Aarseth, E., et al. (2018). A weak scientific basis for gaming disorder: let us err on the side of caution. *J. Behav. Addict.* 7, 1–9. doi: 10.1556/2006.7.2018.19

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* 36, 1–48. doi: 10.18637/jss.v036.i03

Wang, C.-Y., Wu, Y.-C., Su, C.-H., Lin, P.-C., Ko, C.-H., and Yen, J.-Y. (2017). Association between Internet Gaming Disorder and generalized anxiety disorder. *J. Behav. Addict.* 6, 564–571. doi: 10.1556/2006.6.2017.088

Wartberg, L., Kriston, L., Zieglmeier, M., Lincoln, T., and Kammerl, R. (2019). A longitudinal study on psychosocial causes and consequences of Internet gaming disorder in adolescence. *Psychol. Med.* 49, 287–294. doi: 10.1017/S003329171800082X

Wichstrøm, L., Penelo, E., Rensvik Viddal, K., de la Osa, N., and Ezpeleta, L. (2018). Explaining the relationship between temperament and symptoms of psychiatric disorders from preschool to middle childhood: hybrid fixed and random effects models of Norwegian and Spanish children. *J. Child Psychol. Psychiatry* 59, 285–295. doi: 10.1111/jcpp.12772

Wichstrøm, L., Stenseng, F., Belsky, J., von Soest, T., and Hygen, B. W. (2019). Symptoms of internet gaming disorder in youth: predictors and comorbidity. *J. Abnorm. Child Psychol* 47, 71–83. doi: 10.1007/s10802-018-0422-x

Wittek, C. T., Finserås, T. R., Pallesen, S., Mentzoni, R. A., Hanss, D., Griffiths, M. D., et al. (2016). Prevalence and predictors of video game addiction: a study based on a national representative sample of gamers. *Int. J. Ment. Health Addict.* 14, 672–686. doi: 10.1007/s11469-015-9592-8

Wölfling, K., Müller, K. W., Dreier, M., Ruckes, C., Deuster, O., Batra, A., et al. (2019). Efficacy of short-term treatment of internet and computer game addiction: a randomized clinical trial. *JAMA Psychiatry* 76, 1018–1025. doi: 10.1001/jamapsychiatry.2019.1676

World Health Organization [WHO] (2018). *The ICD-11 Classification of Mental and Behavioral Disorders: Diagnostic Criteria for Research*. Available online at: https://icd.who.int/en (accessed August 25, 2021).

Yen, J.-Y., Liu, T.-L., Wang, P.-W., Chen, C.-S., Yen, C.-F., and Ko, C.-H. (2017). Association between Internet Gaming Disorder and adult attention deficit and hyperactivity disorder and their correlates: Impulsivity and hostility. *Addict. Behav.* 64, 308–313. doi: 10.1016/j.addbeh.2016.04.024

Young, K. S. (1998). *Caught in the Net: How to Recognize the Signs of Internet Addiction–and a Winning Strategy for Recovery*. Hoboken, NJ: John Wiley & Sons.

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 January 2022
doi: 10.3389/fpsyg.2021.748672

Deflation-Corrected Estimators of Reliability

*Jari Metsämuuronen**

Finnish National Education Evaluation Centre (FINEEC), Helsinki, Finland

Underestimation of reliability is discussed from the viewpoint of deflation in estimates of reliability caused by artificial systematic technical or mechanical error in the estimates of correlation (MEC). Most traditional estimators of reliability embed product–moment correlation coefficient (PMC) in the form of item–score correlation (*Rit*) or principal component or factor loading (λ_i). PMC is known to be severely affected by several sources of deflation such as the difficulty level of the item and discrepancy of the scales of the variables of interest and, hence, the estimates by *Rit* and λ_i are always deflated in the settings related to estimating reliability. As a short-cut to deflation-corrected estimators of reliability, this article suggests a procedure where *Rit* and λ_i in the estimators of reliability are replaced by alternative estimators of correlation that are less deflated. These estimators are called deflation-corrected estimators of reliability (DCER). Several families of DCERs are proposed and their behavior is studied by using polychoric correlation coefficient, Goodman–Kruskal gamma, and Somers delta as examples of MEC-corrected coefficients of correlation.

Keywords: reliability, deflation in reliability, item-score correlation, deflation in correlation, coefficient alpha, coefficient theta, coefficient omega, maximal reliability

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Ben Kelcey,
University of Cincinnati, United States
Marco Tommasi,
University of Studies G. d'Annunzio Chieti and Pescara, Italy

***Correspondence:**
Jari Metsämuuronen
jari.metsamuuronen@gmail.com

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

Received: *28 July 2021*
Accepted: *15 November 2021*
Published: *04 January 2022*

Citation:
Metsämuuronen J (2022) Deflation-Corrected Estimators of Reliability. Front. Psychol. 12:748672. doi: 10.3389/fpsyg.2021.748672

INTRODUCTION: ATTENUATION AND DEFLATION IN THE ESTIMATES OF RELIABILITY

Reliability of test score (*REL*) is used in several ways of which quantifying the amount of random error in a score variable generated by a compilation of multiple test items may be the most concrete one in the measurement modeling settings. The formula of the average standard error of the measurement $S.E.m. = \sigma_E = \sigma_X\sqrt{1 - REL}$ is derived strictly from the basic definition of reliability $REL = \sigma_T^2/\sigma_X^2 = 1 - \sigma_E^2/\sigma_X^2$, where σ_X^2, σ_T^2, and σ_E^2 refer to the variances of the observed score variable (X) and the unobserved true score (T) and error (E) related to the classic relation of $X = T + E$ (Gulliksen, 1950). Reliability is also used in assessing the (overall) quality of the measurement, in correcting the attenuation of the estimates of regression or path models, in correcting the attenuation in correlations in validity studies and meta-analyses, and for providing confidence intervals around these estimates (see, e.g., Gulliksen, 1950; Schmidt and Hunter, 2015; Revelle and Condon, 2018; Aquirre-Urreta et al., 2019). In all cases, the interest related to the accuracy of the estimates of reliability is understandable.

A less discussed challenge in the estimates by the traditional estimators of reliability is that their estimates may be radically *deflated* caused by artificial systematic errors during the estimation or *attenuated* as a natural consequence of random errors in the measurement (see the discussion of

the terms in, e.g., Chan, 2008; Lavrakas, 2008; Gadermann et al., 2012; Revelle and Condon, 2018); deflation and its correction are the foci in this article. Empirical examples discussed later show that, in certain types of datasets, typically with very easy and very difficult tests and tests with incremental difficulty level including both easy and difficult items, the estimates of reliability may be deflated by 0.40–0.60 units of reliability (see, e.g., Zumbo et al., 2007; Gadermann et al., 2012; Metsämuuronen and Ukkola, 2019; see section "Practical Consequences of Mechanical Error in the Estimates of Correlation in Reliability").

Guttman (1945) was the first to show the technical or mechanical underestimation in the estimators of reliability. He showed that all estimators in his family of estimators λ_1 to λ_6 underestimate the true population reliability. This result generalizes to such known estimators of reliability as Brown–Spearman prophecy formula (ρbs; Brown, 1910; Spearman, 1910), Flanagan–Rulon prophecy formula (ρ_{FR}; Rulon, 1939), coefficient alpha (ρ_α) generalized from Kuder and Richardson (1937) formula KR20 (ρ_{KR20}) by Jackson and Ferguson (1941) and later named by Cronbach (1951), and estimators called the greatest lower bound (ρ_{GLB}; e.g., Jackson and Agunwamba, 1977; Woodhouse and Jackson, 1977) because these are all special cases of $\lambda_1 - \lambda_6$. Hence, using these estimators, the true (population) reliability is always underestimated. Later, Novick and Lewis (1967) pointed out that the underestimation related to the measurement modeling holds if the true values (taus) are not essentially identical and the error components related to the test items do not correlate (see the discussion also in Raykov, 2012; Raykov and Marcoulides, 2017).

Since Guttman (1945), the underestimation in ρ_α has been handled in numerous studies and it has been connected to, among others, a simplified assumption of the classical test theory including unidimensionality, violations in tau–equivalence and latent normality, and uncorrelated errors (see discussion in, e.g., Green and Yang, 2009, 2015; Trizano-Hermosilla and Alvarado, 2016). Some scholars have been ready even to reject ρ_α for all (see, e.g., Yang and Green, 2011; Dunn et al., 2013; Trizano-Hermosilla and Alvarado, 2016; McNeish, 2017) but the discussion is still going on. In many practical testing settings, even though better options are available, ρ_α may still be used as one of the lower bound estimators of reliability because the basic assumptions of alpha such as unidimensionality and uncorrelated errors are usually met (e.g., Metsämuuronen, 2017; Raykov and Marcoulides, 2017).

On the top of attenuation related to the measurement modeling, the estimates of reliability are also deflated—sometimes radically as discussed above. The root cause for the deflation is that the estimates by product-moment correlation coefficient (PMC; Pearson, 1896) embedded in the traditional estimators of reliability in the form of item–score correlation (*Rit*) or principal- or factor loading (λ_i) may be seriously deflated approximating 100% with items with extreme difficulty level and large sample size (see Metsämuuronen, 2020b, 2021b). Deflation in PMC is caused by a phenomenon called here artificial systematic technical or mechanical error in the estimates of correlation (MEC). This phenomenon and its consequences are discussed in section "Mechanical Error in the Estimates of Correlation in PMC and some consequences."

Replacing PMC in the estimators of reliability by a less-MEC-defected coefficient of correlation called later MEC-corrected estimators of correlation leads us to new kinds of estimators of reliability named here *deflation-corrected estimators of reliability* (DCER). DCERs can be divided into two types. One, focused on this article, are MEC-corrected estimators of reliability where PMC is replaced by a totally *different estimator* of correlation that is less prone to deflation than PMC. The other types of DCERs not discussed in this article could be called attenuation-corrected estimators of reliability; in these, PMC is replaced by relevant *attenuation-corrected estimators* of correlation. Some options for the latter are proposed by Metsämuuronen (2021c); attenuation corrected PMC and *eta*. The idea of DCER have been discussed (although not by this name) also, for instance, by Zumbo et al. (2007) and Gadermann et al. (2012) related to their ordinal alpha and ordinal theta; ordinal alpha and theta uses the matrix of inter-item *RPC*s instead of PMCs in the calculations and those are special cases of DCERs.

The crucial role of item–total correlation in the deflation of reliability has been discussed during the years (e.g., Metsämuuronen, 2009, 2016, 2017)[1] and some options of corrected estimators of reliability have been initially suggested, however, without further studies of their behavior (see, e.g., Metsämuuronen and Ukkola, 2019; Metsämuuronen, 2020a,b, 2021b). According to simulations (see, e.g., Metsämuuronen, 2020b, 2021b,d), some good alternatives for PMC are polychoric correlation coefficient (*RPC*; Pearson, 1900, 1913), Goodman–Kruskal gamma (*G*; Goodman and Kruskal, 1954), Somers delta (*D*; Somers, 1962), dimension-corrected *G* and *D* (G_2 and D_2; Metsämuuronen, 2020a, 2021b) and bi- and polyreg correlation (see Livingston and Dorans, 2004; Moses, 2017). Notably, first, some estimators of item–score correlation may be found equally good alternatives or even better than *RPC*, *G*, or *D*. Second, although it seems that nonparametric coefficients of correlation based on order of the cases would be the best options for PMC, this is not categorically true. Of nonparametric options, Kendall's tau-a (Kendall, 1938) and tau-b (Kendall, 1948), as examples, tend to underestimate true correlation even more than PMC (see Kendall, 1949; Metsämuuronen, 2021d; see **Figure 1**).

This article discusses the mechanisms of how the deflation related to coefficients of correlation causes deflation in the estimates of reliability and proposes several concrete options to solve the problem. Numerical examples are given of their behavior. It is asked, what is the effect of changing an estimator with a high quantity of deflation with an estimator with remarkably less deflation in the estimates of reliability? Section "Mechanical Error in the Estimates of Correlation in Product–Moment Correlation Coefficient and Some Consequences" discusses PMC as the root cause of the deflation in reliability, section "Deflation-Corrected Estimators of Reliability" discusses the conceptual base of the DCERs, and sections "Materials and

[1]The basic contents of the derivation of underestimation of PMC in the measurement modeling settings, later elaborated in Metsämuuronen (2016), were initially published in Metsämuuronen (2009); in Finnish.

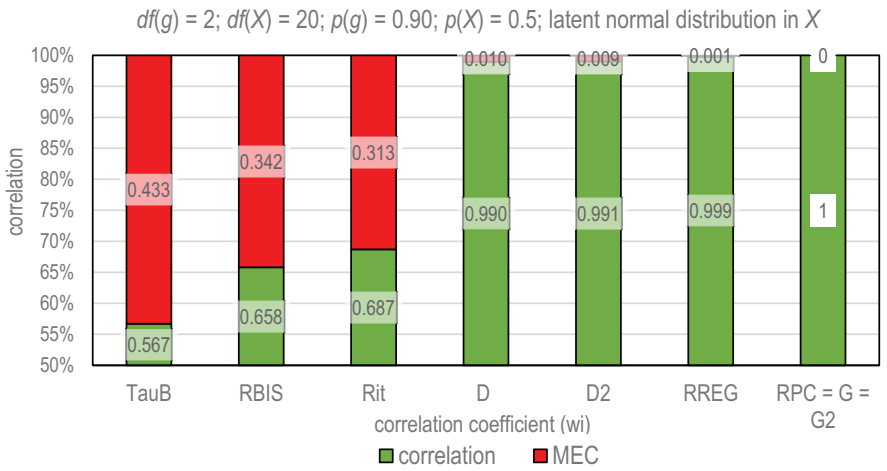

FIGURE 1 | Magnitude of deflation in different estimators. TauB, Kendall tau-b; Rit, PMC; RBIS, biserial correlation; D, Somers delta (X dependent); D2, dimension-corrected D; RREG, r-polyreg correlation; RPC, polychoric correlation; G, Goodman-Kruskal gamma; G2, dimension-corrected G.

Methods" and "Results" give numerical examples of how the deflation in the estimates of reliability is reduced when using DCERs instead of the traditional estimators.

MECHANICAL ERROR IN THE ESTIMATES OF CORRELATION IN PRODUCT–MOMENT CORRELATION COEFFICIENT AND SOME CONSEQUENCES

In measurement modeling settings, MEC refers to a characteristic of estimators of correlation to underestimate the true correlation between the test items (g_i) and the latent trait θ manifested as a score variable (X) caused by artificial technical or mechanical reasons. In what follows, section "Product–Moment Correlation Coefficient, Mechanical Error in the Estimates of Correlation, and Deflation" discusses the overall effect of MEC in PMC, section "Sources of Mechanical Error in the Estimates of Correlation Affecting Deflation in Product–Moment Correlation Coefficient" discusses sources of MEC affecting deflation, section "Product–Moment Correlation Coefficient and the Estimators of Reliability" discusses how PMC is embedded in the estimators of reliability, and section "Practical Consequences of Mechanical Error in the Estimates of Correlation in Reliability" discusses what the effect of deflation in PMC in the estimates of reliability in the empirical dataset may be.

Product–Moment Correlation Coefficient, Mechanical Error in the Estimates of Correlation, and Deflation

The phenomenon of attenuation in the estimates by PMC is well-known. Pearson (1903) and Spearman (1904) may be the first scholars discussing the mechanical errors in estimators of correlation, while Brown (1910) and Spearman (1910) may be the first to connect this to reliability. All of them tried to find a solution to the known challenge in the estimates of correlation known today as restriction of range (see the literature in Sackett and Yang, 2000; Sackett et al., 2007; Meade, 2010). It is known that when only a portion of the range of values of the variable is actualized in a sample it leads to inaccuracy in the estimates of PMC, that is, the values are attenuated. Schmidt and Hunter (1999), specifically, discusses the need of utilizing the knowledge from attenuation correction when estimating measurement error.

Even if there was no obvious restriction of range obtained due to a reduced variance in the score variable within the sample, PMC underestimates the true correlation always if the scales of the variables are not equal (see algebraic reasons in, e.g., Metsämuuronen, 2017). This kind of deflation in PMC caused by mechanical reasons is easy to illustrate by two identical continuous variables with an obvious perfect correlation, $\rho_{XX} = 1$. If we dichotomize one to be a binary variable (item g) and polytomize the other to include several ordinal or interval-scaled bins (score X), PCM between these variables cannot reach the obvious true (perfect latent) correlation. Instead, the value depends, among others, on the cut-off where the ordered continuous variable is dichotomized to 0s and 1s, that is, of the item difficulty. If the cut-off is extreme, PMC approximates 0 irrespective of the fact that the true correlation between the variables was perfect (see simulation e.g., in Metsämuuronen, 2021b). Even at the highest, PMC cannot reach the perfect $\rho_{XX} = 1$; if there are no ties in the score, the highest value approximates 0.866.[2] Then, because of deflation, the loss of information in PMC may vary 13–100% depending on the item difficulty and the sample size. This loss of information is illustrated in **Figure 1**.

To give a practical illustration of the magnitude of error caused by deflation of correlation by different estimators, let us

[2]The value depends on, to some extent, the number of bins in the variable with wider scale. For example, with 10, 20, 30, 200, and 1,000 bins, the maximum value is 0.8704, 0.8671, 0.8665, 0.8660, and 0.8660, respectively. This is easy to confirm by forming these sets of variables.

consider the situation described above: two identical variables with (obvious) perfect correlation $\rho_{XX} = 1$. Let there be 1000 cases and a normal distribution in the original variables. One of the variables becomes an item g by categorizing it into three categories (0, 1, and 2; $df(g) = 2$) and the other is polytomized into 21 categories (score X, $df(X) = 20$). The cut points are arbitrary from the illustration viewpoint; let the average difficulty level of the item be $p(g) = 0.90$ (or, $p(g) = 0.10$) that is, we have a very easy (or difficult) item, and the test score be of a medium difficulty level, $p(X) = 0.50$. **Figure 1** illustrates the differences between some known estimators of correlation; the estimators are discussed later with literature.

Knowing that the latent correlation is perfect, the magnitude of the correlation strictly indicates the amount of deflation. We note that, of the estimators in the example, *tau-b*, biserial correlation (Pearson, 1909), and PMC (*Rit*) cannot reach the (obvious) perfect correlation between the two versions of the same variable and, more, the magnitude of deflation is remarkable (0.43, 0.34, and 0.31 units of correlation, respectively). Of the estimators, D, D_2, and *RREG* give far better approximations of the latent correlation even if there still is some error in the estimates (0.010, 0.009, and 0.001 units of correlation, respectively). In contrast, *RPC*, G, and G_2 reach the perfect latent correlation, that is, there is *no* deflation in the estimates when it comes to *difficulty level* of the items. Notably though, there may be other factors causing deflation or underestimation of association. Some of these factors are discussed in what follows (see also Metsämuuronen, 2021d).

Sources of Mechanical Error in the Estimates of Correlation Affecting Deflation in Product–Moment Correlation Coefficient

By modifying the above example of two identical variables with relevant traditional coefficients of correlations such as *RPC*, G, and D, Metsämuuronen (2021b) concluded that PMC is affected (at least) by six sources of MEC: (1) *Discrepancy in scales of the variables in general*: PMC cannot reach the true (perfect) correlation between the item and the score when the dimensions of the variables differ from each other; (2) *Item difficulty and item variance*: the more extreme the item difficulty, the less variance, and the more underestimation in PMC. The loss of information approximates 100% with extremely easy and difficult items; (3) *The number of categories in the item*: the fewer the categories, the more underestimation in PMC; (4) *The number of categories in the score*: the fewer the categories, the lesser predictable the underestimation is; (5) *The number of tied cases in the score*: more there are tied cases in the score, lesser predictable the underestimation is. This is related to the sample size and the number of categories in the score (point 4); (6) *The distribution of the latent variable*: PMC underestimates the true correlation more if the latent variable is normal or skewed than in the cases of even distribution. These sources of the MEC are not the only possible ones although they are characteristics to PMC (see Metsämuuronen, 2021b).

Although rigorous studies have been done on these elements (e.g., Martin, 1973, 1978; Olsson, 1980; Anselmi et al., 2019; Metsämuuronen, 2021b) these tend to be fragmentary; systematic studies of the several elements of MEC would enrich our knowledge of the phenomenon. Notably, in all the six conditions above related to the attenuation in PMC, such benchmarking coefficients as *RPC* and G appeared to be MEC-free in the simulation (see Metsämuuronen, 2021b); the estimates reach the perfect correlation either strictly ($G = 1$) or asymptotically ($RPC \approx 1$) irrespective of the condition. D appeared to be less affected by MEC than PMC but not to the extent as *RPC* and G (see also **Figure 1**). The reason for the latter is that while *RPC* and G are not affected by the tied cases, D is, specifically, with short tests (see the differences of D and G also in Metsämuuronen, 2021a).

Product–Moment Correlation Coefficient and the Estimators of Reliability

PMC is deep-rooted to the practices within the test theory and measurement modeling settings. From the reliability viewpoint, on the one hand, PMC is *strictly visible* in such classic estimators as ρ_{BS}, ρ_{FR}, ρ_{KR21}, ρ_{α}, ρ_{GLB}, and $\lambda_1 - \lambda_6$ discussed above. Common to these estimators is that the variance of the test score (σ_X^2) inherited from the basic definition of reliability is visible in the formula[3] and σ_X^2, on its behalf, can be expressed by using the item–score correlation ($Rit = \rho_{iX} =$ PMC): $\sigma_X^2 = \left(\sum_{i=1}^{k} \sigma_i \times \rho_{iX}\right)^2$ (Lord et al., 1968) where k refers to number of items in the compilation and σ_i to the standard deviations of partitions or items. Then, as an example, coefficient alpha can be expressed as (Lord et al., 1968):

$$\rho_{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\sigma_X^2}\right) = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times \rho_{iX}\right)^2}\right) \quad (1)$$

On the other hand, PMC is *embedded* in the estimators based on factor- and principal component analysis because the factor- and principal component loadings (λ_i) are, essentially, correlations between an item and the score variable (e.g., Cramer and Howitt, 2004; Yang, 2010). This concerns such estimators of reliability as coefficient theta (ρ_{TH}; Armor, 1973; see also Lord, 1958; Kaiser and Caffrey, 1965), known also as Armor's theta:

$$\rho_{TH} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k}\lambda_i^2}\right), \quad (2)$$

where λ_i are principal component loadings of the (first or only) principal component, coefficient omega (ρ_{ω};

[3]We recall that, although the traditional formula of ρ_{BS} is usually expressed by using PMC between two parallel tests, it can be expressed also by using σ_X^2 in the form familiar from ρ_{FR} (see Lord et al., 1968).

Heise and Bohrnstedt, 1970; McDonald, 1970), known also as McDonald's omega total:

$$\rho_{\omega} = \frac{\left(\sum_{i=1}^{k} \lambda_i\right)^2}{\left(\sum_{i=1}^{k} \lambda_i\right)^2 + \sum_{i=1}^{k} \left(1 - \lambda_i^2\right)}, \quad (3)$$

and coefficient rho, known also as maximal reliability (ρ_{MAX}) or Raykov's rho (Raykov, 1997a, 2004) based on the conceptualization suggested by Li et al. (1996) and Li (1997):

$$\rho_{MAX} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k} \left(\lambda_i^2/\left(1-\lambda_i^2\right)\right)}} \quad (4)$$

(e.g., Cheng et al., 2012) where λ_i are factor loadings.

From the traditional measurement modeling viewpoint (see, e.g., McDonald, 1999; Revelle and Condon, 2018) the forms in Eqs. (1) to (4) implicitly assume that $\rho_{iX}$and λ_i are deflation-free. However, on the one hand, ρ_{iX} is known to be severely deflated (see above). On the other hand, if we use the operationalization familiar in principal component analysis (PCA), exploratory factor analysis (EFA), and structural equation modeling (SEM) where λ_i is a principal component- or factor loading, assumption of deflation-free estimates is too optimistic assumption because λ_i is, essentially, a correlation between item and the factor (or principal component) score variable (Yang, 2010). That is, λ_i is (essentially) ρ_{iX} being deflated as discussed above.

Practical Consequences of Mechanical Error in the Estimates of Correlation in Reliability

The effect of MEC in deflation in the estimates of reliability may be remarkable. Two empirical examples are given. The first example comes from Gadermann et al. (2012) who report a dataset where, by using ordinal alpha (α_{ORD}; Zumbo et al., 2007), another kind of DCER based on replacing the inter-item matrix of PMCs by a matrix of *RPCs*, the estimate by ρ_{α} was deflated from 0.85 (α_{ORD}) to 0.46 (ρ_{α}), that is, 0.39 units of reliability which equals 46% (=0.85–0.46)/0.85).

Another example comes from a national level testing program of learning outcomes (n = 7,770; Metsämuuronen and Ukkola, 2019) where the preconditions of understanding the instruction language were assessed with a very easy 8-item, 11-point test. It was expected that only students with second language background in the instruction language would make mistakes in the test; of all test takers, 72% gave the full marks. The magnitude of the estimate of reliability by the traditional coefficient alpha was found to be ρ_{α} = 0.25 and by rho ρ_{MAX} = 0.48. By using a DCER based on Somers D where $\rho_{iX}$is replaced by $D\,(i\,|X\,) = D_{iX}$ in the formula of alpha (see later Eq. 23), the magnitude of deflation-corrected alpha was ρ_{α_DiX} = 0.86. Then, the magnitude of the estimate by ρ_{α} was deflated around 0.60 units of reliability (71%) and the estimate by ρ_{MAX} around 0.38 units of reliability (44%). The obvious reason for the remarkably higher estimate by ρ_{α_DiX} is that, in the case of binary items with extreme difficulty level, PMC as well as the factor loadings are severely attenuated while, in the binary case, D is less deflated. In both examples, the deflation in the estimates by the traditional estimators is remarkable. The latter example will be re-analyzed in section "Practical Example of Calculating Deflation-Corrected Estimators of Correlations Discussed in This Article" in details.

DEFLATION-CORRECTED ESTIMATORS OF RELIABILITY

Conceptual Base of the Deflation-Corrected Estimators of Reliability

Suggesting a radically new way of estimating reliability urges in-depth discussion of theoretical foundations of the new approach. However, here, the new concepts are built based on the traditional measurement models (see, e.g., McDonald, 1999; Cheng et al., 2012) which are, however, rethought and reconceptualized to also include the elements of deflation. Some further alternatives to consider for rethinking reliability are discussed in section "Options for Correcting the Deflation in Estimators of Reliability." The effect of deflation is discussed here theoretically only to the extent that makes the notation in deflation-corrected estimators of reliability understandable.

Let w_i be a general weight factor that links the observed values (x_i) of an item g_i with the latent variable θ manifested as a score variable:

$$x_i = w_i\theta + e_i \quad (5)$$

generalized from the traditional one-latent variable model (e.g., McDonald, 1999; Cheng et al., 2012). It is relevant to assume that the weight factor w_i is a coefficient of correlation ($-1 \leq w_i \leq +1$) such as *Rit*, *RPC*, *G*, or *D*, or principal component- or factor loadings (λ_i). Also, the latent variable θ may be manifested as varying types of relevantly formed compilation of items such as a raw score (θ_X), factor score variable (θ_{FA}), principal component score variable (θ_{PC}), a theta score formed by the item response theory (IRT) or Rasch modeling (θ_{IRT}), or a possible non-linear compilation of the items (θ_{NonL}).

Eq. (5) generalizes to the compilation of items as

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} w_i\theta + \sum_{i=1}^{k} e_i, \quad (6)$$

where k is the number of items in the compilation. Eq. (6) corresponds with the classic relation of the observed score (X), true score (T), and error (E) in the classical measurement model, that is, $X = T + E$ discussed above. To visualize the differences between different models, this general (congeneric, one-latent variable) model without considering the elements of deflation is as in **Figure 2A**.

From the correlation viewpoint, knowing that all generally used estimators of correlation give identical estimates of the correlation for original variables and for the standardized

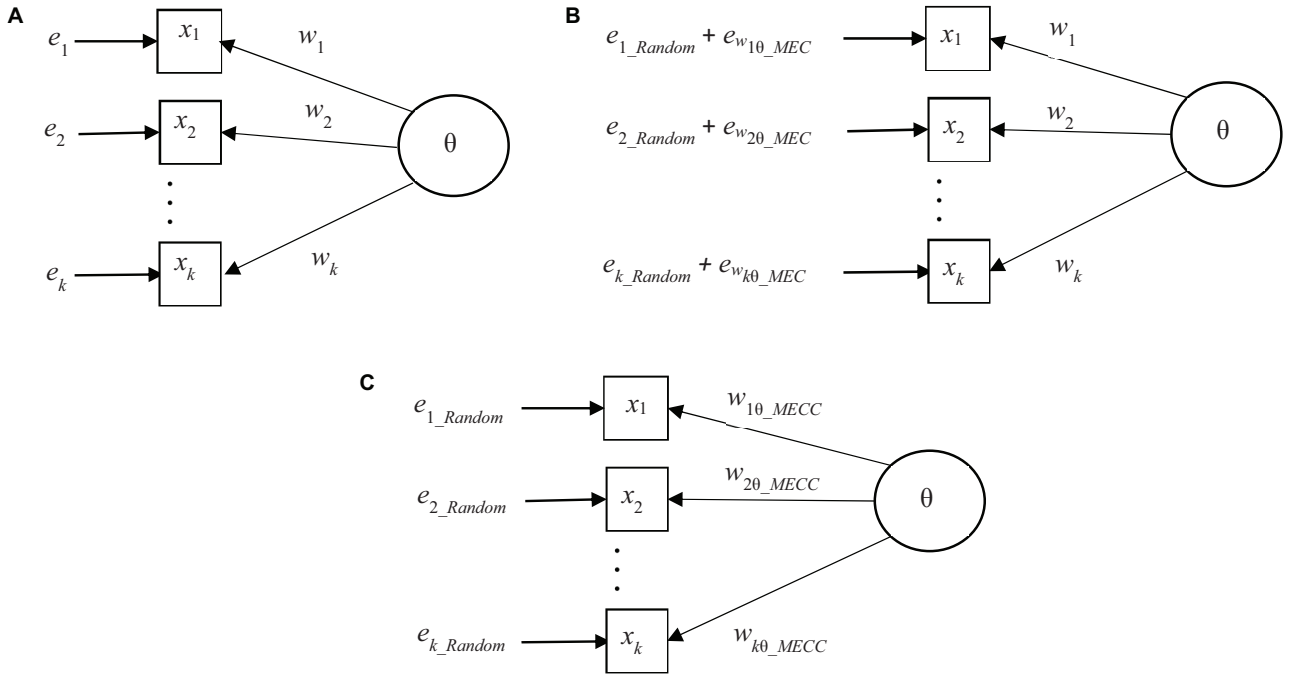

FIGURE 2 | (A) A general one-factor measurement model without elements of related to deflation. **(B)** A general one-factor measurement model with elements of error related to deflation. **(C)** Deflation-corrected one-latent variable measurement model.

versions of the variables, without loss of generality, we can assume that g_i and θ are standardized, $x_i, \theta \sim N(0, 1)$. Then, parallel to the traditional model (see e.g., Cheng et al., 2012), the error variance of the test score ψ_i^2 can be estimated as

$$\psi_i^2 = \sigma_E^2 = VAR(\sum_{i=1}^{k} e_i) = \sum_{i=1}^{k} \left(1 - w_i^2\right). \tag{7}$$

Eq. (7) can be strictly used in estimating the reliability of the score variable ($REL = 1 - \sigma_E^2/\sigma_X^2$). If we use principal component loadings as the weight factor and principal component score as a manifestation of θ, the conceptualization of error variance in Eq. (7) is used strictly in ρ_{TH} (Eq. 2) and, when using factor loadings and factor score variable, it leads to such estimators as ρ_ω and ρ_{MAX} (Eqs. 3 and 4).

The traditional estimators of reliability assume that *Rit* and factor/principal component loadings are deflation-free. This is a too optimistic assumption as discussed and illustrated above (see **Figure 1**). If the observed value of w_i embeds deflation, as it typically does when using the traditional estimators of correlation and loadings, the magnitude of the observed correlation or loading by a deflated or MEC-defected (MECD) weight factor (w_{i_MECD}) is, obviously, lower than MEC-free (MECF) weight factor (w_{i_MECF}), that is,

$$w_{i\_MECF} = w_{i\_MECD} + e_{wi\_MEC} \tag{8a}$$

or

$$w_{i\_MECD} = w_{i\_MECF} - e_{wi\_MEC} \tag{8b}$$

where the exact magnitude of the error element related to deflation in estimation (e_{wi_MEC}) is largely unknown although it is positive ($e_{wi\_MEC} > 0$), and it depends on the characteristics of the item and the weight factor as discussed above. While knowing that a certain part of the measurement error is strictly technical or mechanical in nature, but its magnitude could be reduced, it makes sense to reconceptualize the classic relation of $X = T + E$ into a form

$$X = T + (E_{Random} + E_{MEC}), \tag{9}$$

where the element E_{MEC} related to deflation is something we can deal with. Notably, this kind of "systematic error" is not a kind we usually consider as "systematic" such as a typo in the test item or some technical problem in processes (see Gulliksen, 1950; Krippendorff, 1970). The latter type of error is usually considered harmless from the reliability viewpoint and its effect is added to the random part of the error. Consequently, we can reconceptualize the measurement model in Eq. (5) as

$$x_i = w_i \times \theta + \left(e_{i\_Random} + e_{wi\theta\_MEC}\right), \tag{10}$$

where the notation $e_{wi\theta_MEC}$refers to the fact that the magnitude of the deflation depends on the characteristics of the weighting factor *w*, item *i*, and the score variable θ. This model using a weight factor including radical deflation such as *Rit* or λ_i may be illustrated as in **Figure 2B**. Notably, the magnitude of the total error $\left(e_{i\_Random} + e_{wi\theta\_MEC}\right)$ is, factually, equal to the one seen in the model in **Figure 2A**. However, now the two components are just visual.

While knowing that some estimators of correlation are less deflated than some others, it makes sense to select such coefficient

as the weighting factor where the quantity of technical or mechanical error would be as low as possible. However, it may be difficult to find an estimator of correlation without deflation, that is, that would be totally deflation- or MEC-free. In what follows, the concept of deflation-*corrected* and, specifically, MEC-*corrected* estimator (MECC) is used to refer such estimators where the deflation is known to be radically smaller than in PMC. If selecting wisely the weight factor, the magnitude of error component related to deflation may be near zero, that is, $e_{wi\theta\_MEC} \approx 0$. If we use options of w_i that would lead us to the condition of $e_{wi\theta\_MEC} \approx 0$, because of Eq. (10), this will lead us to a model where the measurement error would be as near the MEC-free condition as possible, that is,

$$\begin{aligned} x_i &= w_{i\_MECC} \times \theta + \left(e_{i\_Random} + e_{wi\theta\_MEC}\right) \\ &\approx w_{i\_MECC} \times \theta + e_{i\_Random}. \end{aligned} \tag{11}$$

This measurement model where MEC-corrected weight factors such as *RPC*, *G*, or *D* are used, could be illustrated as in **Figure 2C**.

As with Eq. (7), knowing that all generally used estimators of correlation give identical estimate of the correlation for original variables (g_i and θ) and for the standardized versions of the variables, we can assume that g_i and θ are standardized, $x_i, \theta \sim N(0, 1)$. Then, assuming that item-wise random errors do not depend on the true scores, the item-wise MEC-corrected error variance ($\psi^2_{i\_MECC}$) is

$$\begin{aligned} \psi^2_{i\_MECC} &= VAR(e_i) = VAR(x_i) - \left(w_{i\_MECC}\right)^2 \times VAR(\theta) \\ &= 1 - w^2_{i\_MECC}, \end{aligned} \tag{12}$$

that is, $e_{i\_MECC} \sim N\left(0, \psi^2_{i\_MECC}\right)$ where $\psi^2_{i\_MECC} = 1 - w^2_{i\_MECC}$. Then, after the deflation-correction, the Eq. (9) could be written as

$$\begin{aligned} X &= T + E_{Random} + E_{MEC} - E_{MEC} \\ &= T + E_{Random} \end{aligned} \tag{13}$$

and Eq. (10) as

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} w_{i\_MECC} \times \theta + \sum_{i=1}^{k} e_{i\_Random}. \tag{14}$$

Consequently, the deflation-corrected error variance of the test score can be written as

$$\sum_{i=1}^{k} \psi^2_{i\_MECC} = \sum_{i=1}^{k} \left(1 - w^2_{i\_MECC}\right), \tag{15}$$

where the form corresponds to the traditional error variance

$$\sum_{i=1}^{k} \psi^2_i = \sum_{i=1}^{k} \left(1 - \lambda^2_i\right) \tag{16}$$

used in the traditional estimators of omega and rho in Eqs. (3) and (4) (see, e.g., Cheng et al., 2012). In the deflation-corrected estimators or reliability, instead of using factor- or principal component loadings we use deflation-corrected estimators of correlation.

Theoretical Deflation-Corrected Estimators of Reliability

By being open for different manifestations of w_i and θ, some options for the base of the deflation-corrected estimators of reliability are theoretical deflation-corrected alpha based on Eq. (1):

$$\rho_{\alpha\_wi\theta} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\left(\sum_{i=1}^{k} \sigma_i \times w_{i\theta}\right)^2}\right), \tag{17}$$

theoretical deflation-corrected theta based on Eq. (2):

$$\rho_{TH\_wi\theta} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} w^2_{i\theta}}\right), \tag{18}$$

theoretical deflation-corrected omega based on Eq. (3):

$$\rho_{\omega\_wi\theta} = \frac{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} w_{i\theta}\right)^2 + \sum_{g=1}^{k}\left(1 - w^2_{i\theta}\right)}, \tag{19}$$

and theoretical deflation-corrected rho based on Eq. (4):

$$\rho_{MAX\_wi\theta} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(w^2_{i\theta}/\left(1-w^2_{i\theta}\right)\right)}}, \tag{20}$$

where $w_{i\theta}$ refers to the general model where the manifestations of θ may vary as well as the linking coefficient w and, obviously, the estimate varies item-wise. Obviously, using the estimators (17) to (20) outside of their original context of raw scores or principal component- and factor analysis is debatable. Here, a stand-point is taken that the forms *could* be used as stand-alone estimators even without their original contexts. This is consistent with a more general measurement model discussed above. Alternatively, the estimators (18) to (20) may be taken as an output of renewed procedures in the principal component- and factor analysis where w_i is a less deflated estimator of correlation than the traditional principal component- and factor loading.

Examples of Practical Deflation-Corrected Estimators of Reliability

By combining the theoretical estimators in Eqs. (17) to (20) and different operationalizations of w_i, we get varying families of deflation-corrected estimator of reliability. Let us assume that we do not fix the manifestation of θ, and we use such MEC-corrected weight factors as *RPC*, *G* and *D* directed so that "item given score" or $D = D(i|X)$ usually labeled as "score dependent" in

the common software packages (of the correct direction of *D*, see Metsämuuronen, 2020b). This leads us to such practical family of deflation-corrected estimators of reliability as deflation-corrected alpha based on Eq. (17) as

$$\rho_{\alpha\_RPCi\theta} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times RPC_{i\theta}\right)^2}\right), \quad (21)$$

$$\rho_{\alpha\_Gi\theta} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times G_{i\theta}\right)^2}\right), \quad (22)$$

and

$$\rho_{\alpha\_Di\theta} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times D\left(g|\theta\right)_{i\theta}\right)^2}\right) = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times D_{i\theta}\right)^2}\right). \quad (23)$$

Because of using totally different type of estimator than PMC, these could be called special types of DCERs, namely, MEC-corrected estimators of reliability. If using some relevant attenuation-corrected estimator of correlation (see some options in Metsämuuronen, 2021c), a family of attenuation-corrected alpha would be obtained.

The notation in names $\rho_{\alpha\_RPCi\theta}$, $\rho_{\alpha\_Gi\theta}$, and $\rho_{\alpha\_Di\theta}$ refers to the facts that the base of the estimator is alpha (α), the weight factor is manifested as *RPC*, *G*, or *D* representing different types of correlations between item and the score variable, and the manifestation of the score variable (θ) could be a raw score (θ_X) or factor score variable (θ_{FA}), as examples. Some of these kinds of estimators are discussed by Metsämuuronen and Ukkola (2019) and Metsämuuronen (2020b, 2021a,b). Another type of solution is discussed by Zumbo et al. (2007) and Gadermann et al. (2012) by replacing the matrix of PMCs by a matrix of *RPC*s in forming the factor loadings; this leads to a coefficient called ordinal alpha discussed above.

More effective estimators than above are expected if coefficient theta (Eq. 18) is used as a base for the estimators and *RPC*, *G*, and *D* as w_i.[4] We get a family of deflation-corrected theta based on Eq. (18):

$$\rho_{TH\_RPCi\theta} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} RPC_{i\theta}^2}\right), \quad (24)$$

$$\rho_{TH\_Gi\theta} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} G_{i\theta}^2}\right), \quad (25)$$

and

$$\rho_{TH\_Di\theta} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} D_{i\theta}^2}\right) \quad (26)$$

or a family of deflation-corrected omega based on Eq. (19):

$$\rho_{\omega\_RPCi\theta} = \frac{\left(\sum_{i=1}^{k} RPC_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} RPC_{i\theta}\right)^2 + \sum_{i=1}^{k}\left(1 - RPC_{i\theta}^2\right)}, \quad (27)$$

$$\rho_{\omega\_Gi\theta} = \frac{\left(\sum_{i=1}^{k} G_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} G_{i\theta}\right)^2 + \sum_{i=1}^{k}\left(1 - G_{i\theta}^2\right)}, \quad (28)$$

and

$$\rho_{\omega\_Di\theta} = \frac{\left(\sum_{i=1}^{k} D_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} D_{i\theta}\right)^2 + \sum_{i=1}^{k}\left(1 - D_{i\theta}^2\right)}, \quad (29)$$

or a family of deflation-corrected rho based on Eq. (20):

$$\rho_{MAX\_RPCi\theta} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(RPC_{i\theta}^2/\left(1-RPC_{i\theta}^2\right)\right)}}, \quad (30)$$

$$\rho_{MAX\_Gi\theta} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(G_{i\theta}^2/\left(1-G_{i\theta}^2\right)\right)}}, \quad (31)$$

[4]The effectiveness is expected because, in their original context, ρ_{TH} maximizes ρ_α (Greene and Carmines, 1980), the magnitude of the estimates by $\rho_{MAX}$is higher than those by ρ_ω (Cheng et al., 2012), and all three give higher value than alpha if the item–score correlations or loadings are not equal (e.g., Cheng et al., 2012).

and

$$\rho_{MAX\_Di\theta} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(D_{i\theta}^2/\left(1-D_{i\theta}^2\right)\right)}}. \quad (32)$$

These families could be called also MEC-corrected theta, omega, and rho. Notably, Zumbo et al. (2007) and Gadermann et al. (2012) also discuss the use of Armor's theta as a basis for ordinal theta by replacing the matrix of PMCs by a matrix of *RPC*s in the estimation.

Many good or even better alternative could be found for *RPC*, *G*, and *D* considering that using *RPC* may lead us to challenges in interpreting the reliability as reflecting unobservable variables (see critique in Chalmers, 2017) and *G* tend to underestimate correlation when there are more than four categories in the item and *D* with three categories or more (see Metsämuuronen, 2021b). For the polytomous case, instead of *G* and *D*, the dimension-corrected *G* and *D* are suggested (Metsämuuronen, 2021b).

The characteristics of the estimators above are not discussed in-depth here; simulations would be beneficial in this matter. However, in the hypothetic extreme datasets with deterministic item discrimination in *all* items leading to $RPC_i = RPC_j \approx G_i = G_j = D_i = D_j = 1$,[5] DCERs based on theta and omega would lead to perfect reliability: $\rho_{TH\_RPCi\theta} \approx \rho_{TH\_Gi\theta} = \rho_{TH\_Di\theta} = k/\left(k-1\right)\left(1-1/k\right) \equiv 1$ and $\rho_{\omega\_RPCi\theta} \approx \rho_{\omega\_Gi\theta} = \rho_{\omega\_Di\theta} = \left(k\right)^2 \Big/ \left(\left(k\right)^2 + 0\right) \equiv 1.$ In the case, estimators (21) to (23) based on alpha can reach the value $\rho_{\alpha\_RPCi\theta} \approx \rho_{\alpha\_Gi\theta} = \rho_{\alpha\_Di\theta} = 1$ only when all item variances are equal ($\sigma_i = \sigma_i = \sigma$), that is, for instance, when the items are standardized. In the case, $\rho_{\alpha\_RPCi\theta} = \rho_{\alpha\_Gi\theta} = \rho_{\alpha\_Di\theta} = k/\left(k-1\right)\left(1-k\sigma^2\Big/\left(k\left(\sigma\times 1\right)\right)^2\right) = k/\left(k-1\right)\times\left(1-1/k\right) \equiv 1$. Otherwise, the maximum value is $\rho_{\alpha\_RPCi\theta}^{Max} \approx \rho_{\alpha\_Gi\theta}^{Max} = \rho_{\alpha\_Di\theta}^{Max} = \frac{k}{k-1}\left(1-\sum_{i=1}^{k}\sigma_i^2\Big/\left(\sum_{i=1}^{k}\sigma_i\right)^2\right)$. Notably, in the deterministic case, estimators based on rho (Eqs. 30 to 32) could not be used because this would require division by zero which is not defined. Aquirre-Urreta et al. (2019) also noted that rho may produce overestimates of the true reliability with finite samples familiar in real-world testing settings. A practical reason for this is that the formula is sensitive to very high values of loadings. In small sample sizes familiar in the real-world datasets, the possibility to obtain deterministic or near-deterministic situation in one or several items increases. In deterministic patterns, ρ_{MAX} cannot be estimated at all and in the near-deterministic patterns the factor loading may be artificially high leading to obvious overestimation in reliability. In what follows in a numerical example, the outcomes based on the DCERs in Eqs. (21) to (23), (30) and (31) are illustrated and the traditional estimators (1) to (4) are used as benchmarks.

[5]Notably, *RPC* cannot reach the perfect 1. With enhanced procedures of the estimation by adding a very small number like 10^{-50} to each element of logarithm and when the embedded PMC ≈ 1 such as 0.99999999, $RPC \approx 1$.

MATERIALS AND METHODS

Dataset Used in the Numerical Example

As a simple numerical example, the dataset consisting of a set of 30 multiple choice questions forming 30 binary items and $n = 49$ randomly selected test-takers from a national level datasets of mathematics test (N = 4,023; FINEEC, 2018) representing small-scale tests with finite samples is used in illustrating the difference between the traditional estimators and deflation-corrected estimators of reliability. The dataset with estimates of different score variables and weight factors are in **Supplementary Appendix 1**.[6]

Measurement Model

The general measurement model discussed in section "Conceptual Base of the Deflation-Corrected Estimators of Reliability" is applied. By using the general one-factor model and by varying w and the operationalization of θ, examples of traditional and deflation-corrected estimates of reliability of the score are given by modifying mainly the form of rho (Eq. 20) with some benchmarking estimates by the form of alpha (Eq. 17).

Operationalizations of the Latent Variable and the Linking Factor

In the empirical section, five operationalizations for θ are used: an unweighted raw score (θ_X), a principal component score variable (θ_{PC}), a factor score variable by maximum likelihood estimation (θ_{FA}), a theta score by one-parameter IRT model or Rasch model (θ_{IRT}), and a nonlinear weighted score by a simple weighting factor $1/p_i$ ($\theta_{NonL} = \theta_{PI} = \sum_{i=1}^{k} g_i/p_i$) where the test-takers are weighted by the proportion of correct answers p_i; the more demanding item, the higher the weight.

Seven options as the weighting factor between θ and g_i are used. First, traditional estimators used in the traditional estimators of reliability: *Rit* with θ_X, principal component loading with θ_{PC}, and ML-estimate of the factor loading with θ_{FA}; second, alternative coefficients *RPC, G,* and *D* for deflation-corrected estimators of reliability; and, third, the traditional PMC (later, *R* or $R_{i\theta}$) as a benchmarking coefficient for the DCERs when not using the traditional alpha. The statistics for and calculations of the estimates are collected in **Supplementary Appendix 1**.

Combining the operationalizations above, we get estimators of reliability related to five different scores and seven linking factors; only selected combinations are used (see condensed in **Table 1**).

[6]The dataset used in this article is a simple one intending to lead the reader to the concepts and relevant estimators by offering all necessary calculations in **Supplementary Appendix 1**. A dataset comprising a more in-depth comparison of different estimators is also available at http://dx.doi.org/10.13140/RG.2.2.27971.94241. This wider dataset is a simulation including 1,440 estimates of reliability drawn from the same real-life dataset as used in **Supplementary Appendix 1**, however, so that the sample size is varied (n = 25, 50, 100, 200) as well as the number of categories and difficulty levels in the items and the score, and more options for the weight element are compared: traditional weights, *RPC*, *G*, *D*, *RREG*, G_2, D_2, *RAC*, and *EAC*. Unlike the dataset used in this article, the score variables in the larger dataset do not include θ_{IRT} and θ_{PI} though.

First, traditional estimators (alpha, theta, omega, and rho; Eqs. 1–4) of which rho is re-notated here to match with the other estimators:

$$\rho_{MAX\_\lambda i\theta_{FA}} = \rho_{MAX} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(\lambda^2_{i\theta_{FA}}\Big/\left(1-\lambda^2_{i\theta_{FA}}\right)\right)}, \quad (33)$$

where the notation $\rho_{MAX\_\lambda i\theta_{FA}}$ refers to facts that coefficient rho is the base of the coefficient (MAX), the manifestation of the score variable is the factor score variable (θ_{FA}), and the manifestation of the weight factor is the ML-estimate of the factor loading ($w_i = \lambda_i = \lambda_{i\theta}FA$).

Second, five estimators based on the form of rho and item–score correlation ($\rho_{i\theta} = R_{i\theta}$) as the linking factor:

$$\rho_{MAX\_Ri\theta_X} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(R^2_{i\theta_X}\Big/\left(1-R^2_{i\theta_X}\right)\right)}, \quad (34)$$

where the score is θ_X and $w_i = R_{i\theta_X}$, (34)

$$\rho_{MAX\_Ri\theta_{PC}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(R^2_{i\theta_{PC}}\Big/\left(1-R^2_{i\theta_{PC}}\right)\right)}, \quad (35)$$

where the score is θ_{PC} and $w_i = R_{i\theta_{PC}}$,

$$\rho_{MAX\_Ri\theta_{FA}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(R^2_{i\theta_{FA}}\Big/\left(1-R^2_{i\theta_{FA}}\right)\right)}, \quad (36)$$

where the score is θ_{FA} and $w_i = R_{i\theta_{FA}}$,

$$\rho_{MAX\_Ri\theta_{IRT}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(R^2_{i\theta_{IRT}}\Big/\left(1-R^2_{i\theta_{IRT}}\right)\right)}, \quad (37)$$

where the score is θ_{IRT} and $w_i = R_{i\theta_{IRT}}$, and

$$\rho_{MAX\_Ri\theta_{PI}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(R^2_{i\theta_{PI}}\Big/\left(1-R^2_{i\theta_{PI}}\right)\right)}, \quad (38)$$

where the score is θ_{PI} and $w_i = R_{i\theta_{PI}}$.

Third, the parallel estimators using $RPC = RPC_{i\theta}$ as the linking factor:

$$\rho_{MAX\_RPCi\theta_X} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(RPC^2_{i\theta_X}\Big/\left(1-RPC^2_{i\theta_X}\right)\right)}, \quad (39)$$

where the score is θ_X and $w_i = RPC_{i\theta_X}$,

$$\rho_{MAX\_RPCi\theta_{PC}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(RPC^2_{i\theta_{PC}}\Big/\left(1-RPC^2_{i\theta_{PC}}\right)\right)}, \quad (40)$$

where the score is θ_{PC} and $w_i = RPC_{i\theta_{PC}}$,

$$\rho_{MAX\_RPCi\theta_{FA}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(RPC^2_{i\theta_{FA}}\Big/\left(1-RPC^2_{i\theta_{FA}}\right)\right)}, \quad (41)$$

where the score is θ_{FA} and $w_i = RPC_{i\theta_{FA}}$,

$$\rho_{MAX\_RPCi\theta_{IRT}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(RPC^2_{i\theta_{IRT}}\Big/\left(1-RPC^2_{i\theta_{IRT}}\right)\right)}, \quad (42)$$

where the score is θ_{IRT} and $w_i = RPC_{i\theta_{IRT}}$, and

$$\rho_{MAX\_RPCi\theta_{PI}} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(RPC^2_{i\theta_{PI}}\Big/\left(1-RPC^2_{i\theta_{PI}}\right)\right)}, \quad (43)$$

where the score is θ_{PI} and $w_i = RPC_{i\theta_{PI}}$.

Fourth, the parallel estimators using $G = G_{i\theta}$ as the linking factor:

$$\rho_{MAX\_Gi\theta_X} = \frac{1}{1+1\Big/\sum_{i=1}^{k}\left(G^2_{i\theta_X}\Big/\left(1-G^2_{i\theta_X}\right)\right)}, \quad (44)$$

TABLE 1 | Estimators of reliability covered in the empirical section.

		Weight factor (the base of the estimator)									
		***Rit(alpha)*[a]**	***RPC(alpha)*[b]**	***G(alpha)*[b]**	***D(alpha)*[b]**	**λ_{PC}*(theta)*[a]**	**λ_{ML}*(omega)*[a]**	**λ_{ML}*(rho)*[a]**	***R(rho)*[b]**	***RPC(rho)*[b]**	***G(rho)*[b]**
	Eqs.	**1**	**21**	**22**	**23**	**2**	**3**	**4, 33**	**34–38**	**39–43**	**44–48**
Score type	θ_X	x	x	x	x				x	x	X
	θ_{PC}					x			x	x	X
	θ_{FA}						x	x	x	x	X
	θ_{IRT}								x	x	X
	θ_{PI}								x	x	X

[a] *Traditional estimates.*
[b] *Deflation-corrected estimates.*

where the score is θ_X and $w_i = G_{i\theta_X}$,

$$\rho_{MAX\_Gi\theta_{PC}} = \frac{1}{1 + 1 \Big/ \sum_{i=1}^{k} \left(G^2_{i\theta_{PC}} \Big/ \left(1 - G^2_{i\theta_{PC}} \right) \right)}, \quad (45)$$

where the score is θ_{PC} and $w_i = G_{i\theta_{PC}}$,

$$\rho_{MAX\_Gi\theta_{FA}} = \frac{1}{1 + 1 \Big/ \sum_{i=1}^{k} \left(G^2_{i\theta_{FA}} \Big/ \left(1 - G^2_{i\theta_{FA}} \right) \right)}, \quad (46)$$

where the score is θ_{FA} and $w_i = G_{i\theta_{FA}}$,

$$\rho_{MAX\_Gi\theta_{IRT}} = \frac{1}{1 + 1 \Big/ \sum_{i=1}^{k} \left(G^2_{i\theta_{IRT}} \Big/ \left(1 - G^2_{i\theta_{IRT}} \right) \right)}, \quad (47)$$

where the score is θ_{IRT} and $w_i = G_{i\theta_{IRT}}$,
and

$$\rho_{MAX\_Gi\theta_{PI}} = \frac{1}{1 + 1 \Big/ \sum_{i=1}^{k} \left(G^2_{i\theta_{PI}} \Big/ \left(1 - G^2_{i\theta_{PI}} \right) \right)}, \quad (48)$$

where the score is θ_{PI} and $w_i = G_{i\theta_{PI}}$.

Additionally, DCERs based on coefficient alpha (Eqs. 21–23) are used as benchmarks to the traditional estimators (see **Table 1**). Of the calculation of the estimates, see **Supplementary Appendix 1**.

RESULTS

Eight outcomes of the comparison are worth highlighting. First, of the estimators based on the form of rho (Eqs. 33 to 48), the ones using *RPC* and *G* as the linking factor give notably higher estimates (0.961–0.968) in comparison to those using PMC (0.894–0.909) and traditional factor- or principal component loadings ($\rho_{MAX} = 0.894$, $\rho_{\omega} = 0.864$, $\rho_{TH} = 0.879$) or alpha ($\rho_{\alpha} = 0.862$) (**Table 2**). This is caused by the better behavior of *RPC* and *G* in relation to deflation with the items with extreme difficulty levels in comparison to PMC (see **Figure 3**). The estimates of reliability based on *RPC* and *G* tend to be more deflation-free than those based on traditional principal component- and factor loadings or PMC, that is, e_{Rit_MEC}, $e_{\lambda i\_MEC} > > e_{RPCi\theta\_MEC} \approx e_{Gi\theta\_MEC}$. The possible overestimation by DCERs is discussed later.

Second, in comparison to the estimates by Eqs. (34) to (38) related to PMC (0.894–0.909) and the traditional ρ_{MAX} (0.894), the estimates by Eqs. (39) to (48) related to *RPC* and *G* tend to be close to each other (0.961–0.969) even though they indicate different aspects of the correlation. While *RPC* estimates the inferred correlation of the (unobservable) latent variables, G estimates the probability that the test takers are in the same order both in an item and a score. The same magnitude of the estimates may be interpreted to indicate that the estimators reflect the same deflation-free reliability of the test score.

Third, the magnitudes of the estimates by the traditional coefficients rho by Eq. (4)($\rho_{MAX\_\lambda i\theta_{FA}} = \rho_{MAX} = 0.894$), theta by Eq. (2) ($\rho_{TH} = 0.879$), and omega by Eq. (3) ($\rho_{\omega} = 0.864$) are higher than by the traditional coefficient alpha by Eq. (1) ($\rho_{\alpha} = 0.862$). This is expected because only in the theoretical case that all the factor loadings or item–score correlations are equal, the magnitude of the estimates by ρ_{α} would reach those by the other estimates. However, it seems that ρ_{MAX} does not produce the "maximal" reliability *per se* for the given test. In the dataset at hand, even the traditional PMC between an item and the factor score variable would lead to a somewhat higher estimate ($\rho_{MAX\_Ri\theta_{FA}} = 0.909$) than using the factor loadings nothing to say of the deflation-corrected estimates ($\rho_{MAX\_RPCi\theta_{FA}} = 0.969$ and $\rho_{MAX\_Gi\theta_{FA}} = 0.968$). Hence, the thinking that "maximal reliability (in the form seen in Eq. 4) is the highest possible reliability that a test can achieve" (Cheng et al., 2012, p. 53 as an example), seems not be true in the absolute sense. Notably though, when using PMC and *RPC* as the linking factor, the score formed by the factor modeling, traditionally taken as the "optimal linear combination" of the items (see, Li, 1997), tends to have the highest reliability in comparison to the other types of score variables although the difference is not notable.

Fourth, coefficient alpha is known to underestimate the true reliability. By using the DCERs based on alpha, that is, Eqs. (21) to (23), the estimates are notably higher ($\rho_{\alpha\_RPCi\theta_X} = 0.937$,

TABLE 2 | Comparison of the estimates of reliability.

		Weight factor (the base of the estimator)									
		Rit(alpha)[a]	*RPC(alpha)*[b]	*G(alpha)*[b]	*D(alpha)*[b]	λ_{PC}*(theta)*[a]	λ_{ML}*(omega)*[a]	λ_{ML}*(rho)*[a]	*R(rho)*[b]	*RPC(rho)*[b]	*G(rho)*[b]
	Eqs.	**1**	**21**	**22**	**23**	**2**	**3**	**4, 33**	**34–38**	**39–43**	**44–48**
Score type	θ_X	0.8619	0.9374	0.9420	0.9343				0.9024	0.9628	0.9682
	θ_{PC}					0.8789			0.9069	0.9661	0.9656
	θ_{FA}						0.8641	0.8943	0.9094	0.9688	0.9681
	θ_{IRT}								0.8944	0.9628	0.9682
	θ_{PI}								0.8987	0.9614	0.9609

[a] *Traditional estimates.*
[b] *Deflation-corrected estimates.*

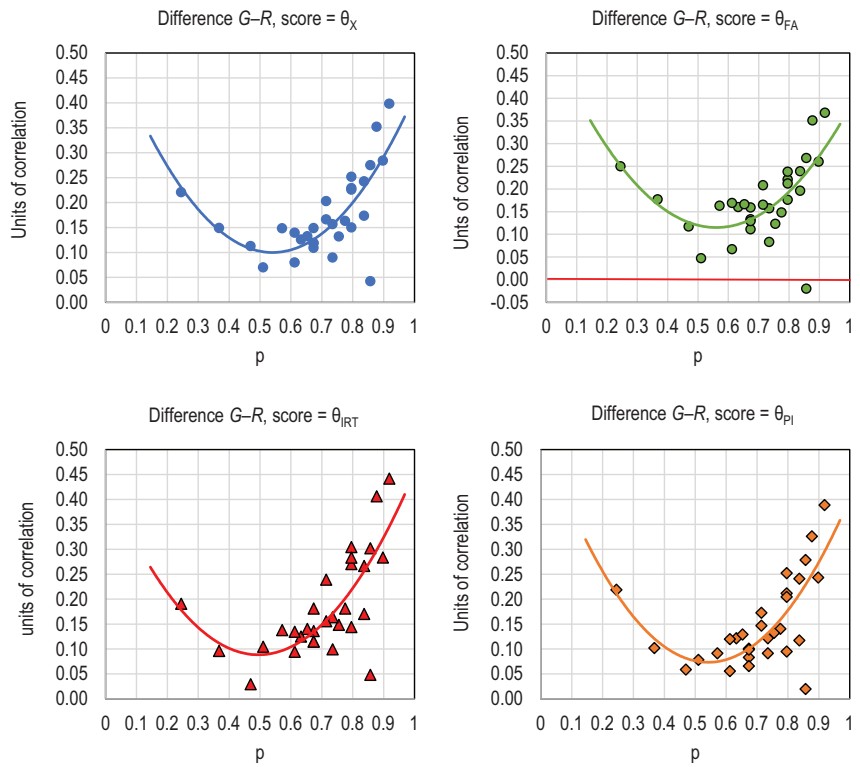

FIGURE 3 | Difference between the estimates of item–score correlation by G and PMC (R).

$\rho_{\alpha\_Gi\theta_X} = 0.942$, and $\rho_{\alpha\_Di\theta_X} = 0.934$), and these are not far from the estimates by the DCERs based on rho with the raw score $\rho_{MAX\_RPCi\theta_X} = 0.963$ by Eq. (39) and $\rho_{MAX\_Gi\theta_X} = 0.968$ by Eq. (44). This seems to indicate that the reliability of the raw score may be closer than what we have thought to the ones manifested as the optimal linear combination of the items.

Fifth, obviously, the outcomes of forming the score differ radically from each other. On the one hand, the scores formed by PCA, EFA, and IRT modeling follow the standardized normal distribution while the raw score and the non-linearly weighted score differ from this logic. On the other hand, the score variables by PCA (θ_{PC}), EFA (θ_{FA}), and non-linear summing (θ_{PI}) do not include tied cases in the dataset; each test takers got their own category in θ_{PC}, θ_{FA} and θ_{PI} while the scores by IRT (θ_{IRT}) and the raw score (θ_X) have identical number of tied cases; in the one-parameter model used in the analysis, θ_{IRT} is a logistic transformation of θ_X. Consequently, the DCERs for the raw score (Eqs. 39 and 44) and for the IRT score (Eqs. 42 and 47) are identical ($\rho_{MAX\_RPCi\theta_X} = \rho_{MAX\_RPCi\theta_{IRT}} = 0.963$ and $\rho_{MAX\_Gi\theta_X} = \rho_{MAX\_Gi\theta_{IRT}} = 0.968$) because the order of the test takers remains the same in the logistic transformation. Regardless of the differences in the structure of the score variables, the estimators based on *G* as a linking factor produce estimates that are largely at the same magnitude of reliability with the scores by raw score, EFA, and IRT by Eqs. (44), (46), and (47): $\rho_{MAX\_Gi\theta_X} \approx \rho_{MAX\_Gi\theta_{FA}} \approx \rho_{MAX\_Gi\theta_{IRT}} \approx 0.968$ and the differences are not wide either when using *RPC* (0.963–0.969). Notably, when using *RPC* and *G* as the linking factor, the score formed by EFA with no tied cases cannot discriminate the test-takers remarkably more accurately than the score with tied cases (θ_{IRT} or θ_X). This reflects the non-obvious fact that reliability of the score variable, in a sense of discriminating the test takers from each other, is not strictly connected with the number of tied values in the score variable nor the type of scale.

Sixth, the obvious reason for the higher magnitude of the estimates by DCERs using *RPC* and *G* in comparison to PMC is caused by the better behavior of *RPC* and *G* with items with extreme difficulty levels. With these kinds of items, specifically, PMC is highly deflated while *RPC* and *G* are not at all affected by item difficulty (see simulation in Metsämuuronen, 2021b). The difference between the estimates of correlation by PMC and *G* is illustrated in **Figure 3**; the graphs would be essentially identical with PMC and *RPC* because the difference between the estimates by *RPC* and *G* are subtle in binary case (see Metsämuuronen, 2020b, 2021b).

Seventh, Green and Yang (2009) approximate that, by using ρ_α, the true reliability may be underestimated up to 11% although, in real-life testing settings, the underestimation may be nominal (Raykov, 1997b). Assuming that *RPC* does not overestimate correlation, when knowing the magnitude of the estimate by the traditional coefficient alpha related to the raw score by Eq. (1) ($\rho_\alpha = 0.862$) and the deflation-corrected estimate by *RPC* related to the factor score variable by Eq. (33)

($\rho_{MAX\_RPCi\theta_{FA}} = 0.969$) in the given dataset, the magnitude of the deflation in the traditional estimate by ρ_α appears to be 0.1068 units of reliability, that is, 11.0% (=(0.969–0.862)/0.969) in comparison to the one by deflation-corrected rho. By using the same logic, the traditional maximal reliability $\rho_{MAX} = 0.894$ is deflated by 7.7%. These seem decent magnitudes considering that, in the empirical cases, the deflation may be 70 or 44% as discussed in section "Practical Consequences of Mechanical Error in the Estimates of Correlation in Reliability." The reason for the decent deflation is that the dataset used in the example is neither extremely easy nor extremely difficult. An obvious confounding factor is that the score variables differ between coefficients alpha and rho. If the score variable would be harmonized as being the raw score and the weighting factor would be harmonized to *RPC*, we can assess the pure effect of the estimator itself. The magnitude of the deflation-corrected alpha (Eq. 21) is ρ_{α_RPCiX} = 0.937 and the magnitude of the deflation-corrected rho (Eq. 39) is ρ_{MAX_RPCiX} = 0.963. Then, the deflation would be reduced from 11 to 2.6% (=(0.963–0.937)/0.963). This (around) 3% seems to refer strictly to a more effective estimation of reliability by using the form of estimator based on maximal reliability than by the formula used in the traditional coefficient alpha. Obviously, more studies are needed to confirm the results.

Finally, eighth, by comparing the estimates of different weighting factors w_i, it is possible to evaluate roughly what the magnitude of the deflation ($e_{wi\theta\_MEC}$) in different estimators of correlation in the dataset is. Assuming that the estimates by *RPC* do not overestimate the correlation between the items and the score, the difference between the estimates based on *RPC* and PMC gives a hint of the magnitude of the deflation in PMC. On average in the given dataset, the deflation in PMC with different types of score variable is $\bar{e}_{Ri\theta_X\_MEC} = 0.156$ units of correlation with raw score (ranging 0.0279–0.3268 depending on the item), $\bar{e}_{Ri\theta_{FA}\_MEC} = 0.157$ (–0.0064–0.3121) with the factor score, $\bar{e}_{Ri\theta_{IRT}\_MEC} = 0.166$ (0.0315–0.3702) with the theta score by IRT modeling, and $\bar{e}_{Ri\theta_{PI}\_MEC} = 0.153$(0.0061–0.3433) with the non-linearly weighted score. The systematic negative bias of this size has a notable effect in deflation in the estimate of reliability.

CONCLUSION AND LIMITATIONS

An obvious conclusion of the theoretical and empirical parts of the study is that the magnitude of the deflation of reliability depends not only on the unidimensionality, violations in the measurement model and latent normality, estimator of reliability, and uncorrelated errors as traditionally suggested with coefficient alpha but also on the estimators of correlation used as the linking factor between the latent trait θ and the test items g_i. Some linking factors like PMC are more prone to deflation than some other estimators like *RPC*, *G* and *D* as examples and, hence, the estimates by PMC are more deflated than those by *RPC*, *G* and *D*. Because PMC is embedded in the traditional estimators of reliability, the deflation in correlation is inherited to the estimates of reliability. Systematic studies comparing different estimators of correlation and reliability could be beneficial to understand the phenomenon better.

Options for Correcting the Deflation in Estimators of Reliability

The root challenge related to deflation in the traditional estimators of reliability seems to be the classical definition of reliability based on *variances* (σ_X^2, σ_T^2, and σ_E^2) leading to use PMC in the practical solutions of estimating reliability. If we would start to create a theory concerning reliability by knowing all the deficiencies of PMC we know today, we may be trying to avoid PMC and, consequently, the variances in the process. To rectify this root challenge, it may be beneficial to rethink the definition of reliability from this perspective. Alternative bases to consider for rethinking reliability may be related to, among other, "sufficiency of information" by Smith (2005), or several options within IRT modeling such as "person separation" by Andrich and Douglas (1977), Andrich (1982), and Wright and Masters (1982), or "information function" discussed by, e.g., McDonald (1999), Cheng et al. (2012), and Milanzi et al. (2015). One alternative for defining reliability is discussed briefly here based on Metsämuuronen (2020b) related to the definition of "ultimately discriminating test score."

Metsämuuronen (2020b) proposes an operational definition of the *ultimate item discrimination* as a condition where the score can predict response pattern of the test-takers in a single item in a deterministic manner. This could be generalized as a theoretical condition for ultimate reliability as being a condition where the score can predict the order (or item response pattern) of the test takers in a deterministic manner *in all items*. This operational definition alone is not very practical when it comes to estimation of the reliability because the deterministic patterns cannot be estimated by using maximum likelihood method, for example. However, this could be a starting point to develop estimators where different types of estimators of item discrimination as well as *a*-parameter in IRT-modeling could be a visible part of the estimator as in Eqs. (21) to (32). Theoretical and empirical work in this area would be beneficial.

While waiting for development of a sound basis for a new way of thinking, defining, and estimating reliability, practical options lead to a kind of new paradigm in the settings related to measurement modeling: the extended families of deflation-corrected estimators of reliability. One set of family, attenuation-corrected estimators of reliability, not discussed in this article, would be obtained if attenuation-corrected estimators of PMC were used instead of PMC in the estimators. Another set of family, MEC-corrected estimators of reliability focused in this article, is obtained if PMC is replaced by a totally different estimator of correlation that would not be deflated at all or where the magnitude of deflation is remarkably smaller than that in PMC. Several new estimators of deflation-corrected estimators were proposed based on using *RPC*, *G* and *D* as examples instead of PMC in some known estimators of reliability.

In the empirical part, it was demonstrated that if *RPC*, *G*, or *D* would be used instead of PMC in some known

estimators of reliability, the deflation in reliability would be corrected to a notable extent. Further simulations with different types of datasets, different item types, different weighting factors, and different base of the estimators (e.g., alpha, theta, omega, or rho) would be beneficial in this regard. The estimates by deflation-corrected estimators are not, factually, "real" reliabilities as such. However, they are *closer* to the deflation-free reliability than the traditional estimates. Empirical examples show that, in specific forms of datasets as in very easy or very difficult tests, the estimates by traditional estimators such as coefficient alpha and rho may be deflated 40–70% because of technical reasons. The DCERs discussed in this article are strong with these kinds of datasets and could be used as a benchmark to the traditional estimators.

Practical Example of Calculating Deflation-Corrected Estimators of Correlations Discussed in This Article

To give a practical example of the DCERs discussed in this article, let us re-analyze the reliability of the extremely easy dataset (n = 7,770) by Metsämuuronen and Ukkola (2019) discussed in section "Practical consequences of Mechanical Error in the Estimates of Correlation in reliability." The advance of DCERs may be notable in these kinds of datasets where the item difficulties are extreme leading to an ultimately non-normal score (see **Table 3**). Because of ultimately easy items with mainly binary scales combined with a non-normal score variable, the non-parametric coefficients of correlation may be better options than PMC.

Deflation-Corrected Alpha

The traditional coefficient alpha uses raw score (θ_X) as the manifestation of the latent ability and item–score correlation (R_{gX}) as the weighting element in the calculation. Estimates by alternative coefficients of item–score association are collected in **Table 4**; their calculation is described in **Supplementary Appendix 1**. Notably, first, the magnitudes of the estimates by *Rit* (0.38 on average) are remarkably lower than those by *RPC* (0.72), *G* (0.88), and *D* (0.83). This is caused by its poor behavior with items of extreme difficulty level. Second, the magnitude of the estimates by *RPC* is somewhat lower than those by *G* and *D*. This is not a general characteristic of these coefficients. With binary items, the estimates by *G* and *RPC* tend to be very close each other (see, e.g., Metsämuuronen, 2021b), and when the number of categories in the item increases up to four or higher, the probability that two variables are in the same order indicated by *G* (and *D*) tend to be lower than covariation between the two variables indicated by *PMC* and *RPC* and, hence, the estimates would signal that the true correlation is underestimated (see Metsämuuronen, 2021b). Third, that the magnitude of the estimates by *D* are lower than those by *G* is expected because the estimates by *D* are more conservative in comparison with *G* (e.g., Metsämuuronen, 2021a,b).

Because of Eq. (1), the traditional coefficient alpha gives the estimate: $\rho_\alpha = \frac{8}{8-1}\left(1 - \frac{0.600}{0.874^2}\right) = 0.245$. The deflation-corrected alpha using *RPC* as the weighting element (Eq. 21)

TABLE 3 | Descriptive statistics of the dataset from Metsämuuronen and Ukkola (2019).

Item (g)	N	Maximum	Mean	*p*	*SD*	Score	Freq.	%
g1	7,770	1	0.96	0.96	0.186	3	4	0.1
g2	7,770	1	0.98	0.98	0.126	4	7	0.1
g3	7,770	1	0.99	0.99	0.088	5	6	0.1
g4	7,770	1	0.91	0.91	0.287	6	20	0.3
g5	7,770	2	1.78	0.89	0.610	7	40	0.5
g6	7,770	1	0.98	0.98	0.122	8	141	1.8
g7	7,770	2	1.97	0.985	0.211	9	809	10.4
g8	7,770	2	1.98	0.99	0.169	10	903	11.6
						11	5,840	75.2
							7,770	100.0

TABLE 4 | Item–score correlations and related statistics needed in estimating reliability.

Item (g_i)	R_{gX}[a]	D_{gX}[a]	G_{gX}[a]	RPC_{gX}[a]	σ_g^2 = VAR(g)	$R_{gX} \times \sigma_g$	$D_{gX} \times \sigma_g$	$G_{gX} \times \sigma_g$	$RPC_{gX} \times \sigma_g$
g1	0.351	0.791	0.857	0.677	0.035	0.065	0.147	0.160	0.126
g2	0.268	0.779	0.846	0.618	0.016	0.034	0.098	0.107	0.078
g3	0.283	0.858	0.911	0.696	0.008	0.025	0.076	0.080	0.061
g4	0.458	0.789	0.834	0.736	0.082	0.131	0.226	0.239	0.211
g5	0.746	0.952	0.979	0.931	0.372	0.455	0.580	0.597	0.568
g6	0.260	0.766	0.831	0.602	0.015	0.032	0.094	0.102	0.074
g7	0.327	0.832	0.897	0.702	0.045	0.069	0.176	0.189	0.148
g8	0.373	0.877	0.924	0.760	0.028	0.063	0.148	0.156	0.128
				SUM	0.600	0.874	1.546	1.630	1.395

[a] *R, Pearson correlation; D, Somers delta "X dependent"; G, Goodman–Kruskal gamma; RPC, polychoric correlation coefficient.*

TABLE 5 | Principal component loadings and related alternative statistics for estimating reliability.

Item (g)	λ_{iPC}	$(\lambda_{iPC})^2$	$D_{g\theta PC}$	$(D_{g\theta PC})^2$	$G_{g\theta PC}$	$(G_{g\theta PC})^2$	$RPC_{g\theta PC}$	$(RPC_{g\theta PC})^2$
g1	0.444	0.197	0.937	0.878	0.937	0.878	0.833	0.694
g2	0.429	0.184	0.960	0.922	0.960	0.922	0.837	0.701
g3	0.593	0.352	0.994	0.988	0.994	0.988	0.947	0.897
g4	0.478	0.228	0.892	0.796	0.892	0.796	0.818	0.669
g5	0.207	0.043	0.737	0.543	0.737	0.543	0.647	0.419
g6	0.375	0.141	0.939	0.882	0.939	0.882	0.791	0.625
g7	0.286	0.082	0.856	0.733	0.856	0.733	0.659	0.435
g8	0.628	0.394	0.984	0.968	0.984	0.968	0.926	0.858
SUM		1.621		6.709		6.709		5.297

TABLE 6A | Factor loadings and related alternative statistics for estimating omega.

Item (g)	λ_i	$(\lambda_i)^2$	$1–(\lambda_i)^2$	$D_{g\theta ML}$	$(D_{g\theta ML})^2$	$1–D_{g\theta ML}{}^2$	$G_{g\theta ML}$	$(G_{g\theta ML})^2$	$1–G^2$	$RPC_{g\theta ML}$	$(RPC_{g\theta ML})^2$	$1–RPC_{g\theta ML}{}^2$
g1	0.276	0.076	0.924	0.940	0.884	0.116	0.940	0.884	0.116	0.831	0.691	0.309
g2	0.260	0.068	0.932	0.957	0.916	0.084	0.957	0.916	0.084	0.829	0.688	0.312
g3	0.471	0.222	0.778	0.995	0.990	0.010	0.995	0.990	0.010	0.962	0.926	0.074
g4	0.291	0.085	0.915	0.892	0.796	0.204	0.892	0.796	0.204	0.814	0.663	0.337
g5	0.111	0.012	0.988	0.736	0.542	0.458	0.736	0.542	0.458	0.645	0.415	0.585
g6	0.213	0.045	0.955	0.934	0.872	0.128	0.934	0.872	0.128	0.774	0.599	0.401
g7	0.160	0.026	0.974	0.844	0.712	0.288	0.844	0.712	0.288	0.660	0.435	0.565
g8	0.512	0.262	0.738	0.993	0.986	0.014	0.993	0.986	0.014	0.960	0.922	0.078
SUM	2.294		7.204	7.291		1.302	7.291		1.302	6.475		2.661

leads to an estimate $\rho_{\alpha\_RPCiX} = \frac{8}{8-1}\left(1 - \frac{0.600}{1.395^2}\right) = 0.790$, gamma (Eq. 22) to $\rho_{\alpha\_GiX} = \frac{8}{8-1}\left(1 - \frac{0.600}{1.630^2}\right) = 0.885$, and delta (Eq. 23) to $\rho_{\alpha\_DiX} = \frac{8}{8-1}\left(1 - \frac{0.600}{1.546^2}\right) = 0.856$. The estimate by the traditional coefficient alpha is radically deflated, 72%, when comparing it to the DCER using *G* as the weighting element $((0.885 - 0.245)/0.885 = 0.723)$ and 69% if using *RPC*. We also note that the magnitude of the estimates of reliability follows strictly the general tendency of the magnitudes of the coefficients of correlation: In comparison with the estimate by ρ_{α_GiX} the estimate by ρ_{α_DiX} is conservative.

Deflation-Corrected Theta

The traditional coefficient theta uses principal component score (θ_{PC}) as the manifestation of the latent ability and principal component loadings (λ_i) as the weighting element in the calculation. Loadings and corresponding statistics related to alternative estimators are collected in **Table 5**. Notably, because there appeared to be no tied pairs between the principal component score and items, the estimates by *G* and *D* are identical.

The traditional coefficient theta can be calculated by Eq. (2): $\rho_{TH} = \rho_{TH\_\lambda i\theta_{PC}} = \frac{8}{8-1}\left(1 - \frac{1}{1.621}\right) = 0.438$. The deflation-corrected theta using *RPC* as the weight factor and the principal component score (θ_{PC}) as the manifestation of the latent ability (Eq. 24) leads us to an estimate $\rho_{TH\_RPCi\theta_{PC}} = \frac{8}{8-1}\left(1 - \frac{1}{5.297}\right) = 0.927$, gamma (Eq. 25) leads to $\rho_{TH\_Gi\theta_{PC}} = \frac{8}{8-1}\left(1 - \frac{1}{6.709}\right) = 0.973$, and delta (Eq. 26) to $\rho_{\alpha\_Di\theta_{PC}} = \frac{8}{8-1}\left(1 - \frac{1}{6.709}\right) = 0.973$. If the estimates based on *G* or *D* are used as a reference value, the traditional coefficient theta is deflated by 54%, and, if *RPC* is used, 52%. If the raw score (θ_X) would be used as a manifestation of the latent ability instead of θ_{PC}, based on the estimates of correlation in **Table 4**, the magnitudes of the latter estimates would be $\rho_{TH\_RPCiX} = 0.869$, $\rho_{TH\_GiX} = 0.961$, and $\rho_{TH\_Di\,X} = 0.937$.

Deflation-Corrected Omega and Rho

The traditional coefficients omega and rho use maximum likelihood estimates of factor score (θ_{ML}) as the manifestation of the latent ability and factor loadings (λ_i) as the weighting element in the calculation. Loadings and corresponding statistics related to alternative estimators are collected in **Tables 6A,B**. As with principal component analysis, because there are no tied

TABLE 6B | Statistics for calculating rho based on **Table 6A**.

Item (g)	$(\lambda_i)^2/(1–(\lambda_i)^2)$	$(D_{g\theta ML})^2/(1–(D_{g\theta ML})^2)$	$(G_{g\theta ML})^2/(1–(G_{g\theta ML})^2)$	$(RPC_{g\theta ML})^2/(1–(RPC_{g\theta ML}{}^2)$
g1	0.082	7.591	7.591	2.232
g2	0.073	10.883	10.883	2.202
g3	0.285	99.251	99.251	12.545
g4	0.093	3.894	3.894	1.971
g5	0.012	1.182	1.182	0.711
g6	0.048	6.834	6.834	1.494
g7	0.026	2.476	2.476	0.771
g8	0.355	70.679	70.679	11.776
SUM	0.974	202.791	202.791	33.701

TABLE 7 | Summary of estimates of reliability.

		Traditional estimate	DCERs with the traditional score			DCERs with the raw score		
Form	**Score type (θ)**	***R***	***D***	***G***	***RPC***	***D***	***G***	***RPC***
Alfa	Raw score (θ_X)	0.245	0.856	0.885	0.790	0.856	0.885	0.790
Theta	Principal component score (θ_{PC})	0.444	0.973	0.973	0.927	0.937	0.961	0.869
Omega	Factor score (θ_{ML})	0.422	0.976	0.976	0.940	0.947	0.967	0.895
Rho	Factor score (θ_{ML})	0.493	0.995	0.995	0.971	0.961	0.979	0.929

pairs between the factor score and items, the estimates by *G* and *D* are identical.

By Eq. (3), the traditional coefficient omega total is calculated as follows: $\rho_{\omega} = \rho_{\omega\_\lambda i\theta_{\mathrm{ML}}} = \frac{(2.294)^2}{(2.294)^2+7.204} = 0.422$ and rho by Eq. (4): $\rho_{MAX} = \rho_{MAX\_\lambda i\theta_{\mathrm{ML}}} = \frac{1}{1+1/0.974} = 0.493$. The deflation-corrected omega using *RPC* as the weight factor (Eq. 27) and the factor score (θ_{ML}) as the manifestation of the latent ability leads us to an estimate $\rho_{\omega\_RPCi\theta_{ML}} = \frac{(6.475)^2}{(6.475)^2+2.661} = 0.940$ and the corresponding deflation-corrected rho (Eq. 30) is $\rho_{MAX\_RPCi\theta_{ML}} = \frac{1}{1+1/33.701} = 0.971$. Similarly, deflation-corrected omega using gamma (Eq. 28) leads to $\rho_{\omega\_Gi\theta_{ML}} = \frac{(7.291)^2}{(7.291)^2+1.302} = 0.976$ and the corresponding deflation-corrected rho (Eq. 31) is $\rho_{MAX\_Gi\theta_{ML}} = \frac{1}{1+1/202.791} = 0.995$. Deflation-corrected omega using delta (Eqs. 29) leads to identical estimates in comparison with the estimates by gamma:$\rho_{\omega\_Di\theta_{ML}} = \frac{(7.291)^2}{(7.291)^2+1.302} = 0.976$ and the corresponding deflation-corrected rho (Eq. 32) is $\rho_{MAX\_Di\theta_{ML}} = \frac{1}{1+1/202.791} = 0.995$.

The magnitude of the estimates based on the form of maximal reliability and *G* and *D* as the weighting factor (0.995), feel intuitively overestimates. This is reasoned by the fact that the formula of maximal reliability is sensitive for high values of loadings. With very high values of loading—as here $G = D = 0.995$ for item *g*3 referring to a fact that after the test takers are ordered by the factor score variable, 99.5% of the test takers are in the same order in both item and score—the statistic $\lambda_i^2/(1-\lambda_i^2)$ may give an artificially high value leading to artificially high estimate of reliability. However, if the estimates based on *G* or *D* are used as a reference value, the traditional coefficient omega and rho are deflated by 57 and 50%, and, if *RPC* is used, 55 and 49%, respectively. If the raw score (*X*) would be used as a manifestation of the latent ability instead of θ_{ML}, the magnitudes of the DCERs based on omega would be $\rho_{\omega\_RPCiX} = 0.895$, $\rho_{\omega\_GiX} = 0.967$, and $\rho_{\omega\_DiX} = 0.947$ and DCERs based on rho $\rho_{MAX\_RPCiX} = 0.929$, $\rho_{MAX\_GiX} = 0.979$, and $\rho_{\omega\_Di\,X} = 0.961$.

The estimates of reliability above are summarized in **Table 7**. Different interpretations of the varying estimators are discussed in the next section. Anyhow, just by comparing the overall level of magnitudes of the traditional estimates and the estimates by different DCERs we may conclude that all the DCERs seem to refer to a reliability which is notably higher than the ones indicated by the traditional estimators. If one uses the raw scores, instead of $\rho_{\alpha} = 0.245$, the true reliability seems to be around 0.914 (on average), varying between 0.790 and 0.979 depending on which form is used as the base and which deflation-corrected estimator of correlation is used as the weighting element. Knowing the interpretation of *RPC*, *G* and *D*, the high magnitude of reliability by DCERs refer to the fact that the score is highly capable of ordering the test takers in a logical order by their latent ability. Of the estimators, the ones based on coefficient alpha are the most conservative and the ones based on rho the most liberal. In this case, the estimators of correlation based on probability (*G* and *D*) tend to lead somewhat higher estimates than the one based on covariance (*RPC*). This is not a general characteristic though.

Different Interpretation of Different Estimators of Reliability

The article did not tackle the issue of differences between the estimators of correlation. Notably, PMC, *RPC*, and *G* (as well as *D*) discussed in the article indicate different aspects of the correlation: PMC estimates the *observed correlation* between two variables, and this is radically deflated in the measurement modeling settings. *RPC* estimates the *inferred correlation* of two unobservable continuous variables by their ordinal manifestations. *G* and *D* estimate the *probability* that the test takers are in the same order both in an item and a score. The outcome of different estimators of reliability may, then, indicate different viewpoints of reliability.

Chalmers (2017) is skeptical of the usefulness of coefficients using *RPC* in practical settings because *RPC* refers to correlation between unobservable and unreachable variables and, therefore, the outcome may be useless in the factual interpretation of the observed score. He proposes that using *RPC* leads to infer something about *theoretical reliability*. However, some estimators of reliability such as ordinal alpha and theta by Zumbo et al. (2007; see also Gadermann et al., 2012), factually, use *RPC* in the estimation. Comparing the estimators related to *RPC* in Eqs. (21), (24), and (27) and (39) to (43) with ordinal alpha or ordinal theta based on the matrix of inter-item *RPCs* instead of matrix of PMCs may be worth studying.

Estimators based on *G* and *D* refer to observed variables and, therefore, the outcome may be more useful than those by *RPC* in the factual analysis of the observed score. Knowing the interpretation of *G* and *D* in the measurement settings (see Metsämuuronen, 2021a,b), estimators (22) and (23), (25) and (26), (31) and (32), and (44) to (48) reflect the average proportion

TABLE 8 | General typological characteristics of selected options of DCERs.

		Weight w_i	
		RPC	***G* and *D***
Base	General characteristics	• Reflects *latent* reliability; not strictly related to the observed score nor observed items • Leads to theoretical interpretation of reliability • Based on covariance • Suitable for binary and polytomous items • Not simple to calculate	• Reflects reliability of the *observed* score • Leads to practical interpretation of reliability • Based on probability • *D* is more conservative than *G* • Suitable for binary items and polytomous items with < 4 categories (*D*) or with < 5 categories (*G*) • Simple to calculate even manually
Alpha	• Always underestimates population reliability • Very conservative nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (*REL* = 1) when $w_i = 1$, and $\sigma_i = \sigma_j$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times RPC_{i\theta}\right)^2}\right)$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i \times G_{i\theta}\right)^2}\right)$
Theta	• Maximizes alpha • Conservative nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (*REL* = 1) when $w_i = 1$	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}RPC_{i\theta}^2}\right)$	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}G_{i\theta}^2}\right)$
Omega	• Estimates always higher than alpha • Least conservative nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (*REL* = 1) when $w_i = 1$	$\frac{\left(\sum_{i=1}^{k}RPC_{i\theta}\right)^2}{\left(\sum_{i=1}^{k}RPC_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-RPC_{i\theta}^2\right)}$	$\frac{\left(\sum_{i=1}^{k}G_{i\theta}\right)^2}{\left(\sum_{i=1}^{k}G_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-G_{i\theta}^2\right)}$
Rho (maximal reliability)	• Maximizes omega • Liberal nature; may overestimate reliability with small sample sizes • Cannot be calculated if deterministic patterns ($\lambda = 1$) even in one item • Cannot reach the perfect reliability (*REL* = 1) • Not the best option for small samples	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(RPC_{i\theta}^2/\left(1-RPC_{i\theta}^2\right)\right)}}$	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(G_{i\theta}^2/\left(1-G_{i\theta}^2\right)\right)}}$

of logically ordered test takers in all items as a whole. In this, the estimators based on *D* are more conservative than the ones based on *G*.

A relevant question is, how different is the interpretation of the estimates by *G* (or *D*) in comparison to those by PMC or *RPC*? Knowing that *G* estimates the probability that the test takers are in the same order in the item and in the score, the ultimate magnitude of reliability by the estimators based on *G* would indicate that *all* items discriminate the higher-performing test takers from the lower-performing test takers in a deterministic manner after the test takers are ordered by the score. The same interpretation would be obtained when using *RPC* except that *RPC* can reach the value *RPC* = 1 only approximatively. From this viewpoint, the deflation-corrected estimators in Eqs. (24) to (32) related to *RPC*, *G*, and *D* seems to refer strictly to the *discrimination power* of the score. This makes sense from the standard error of measurement viewpoint. Notably, under the condition of deterministic item discrimination, the estimators using PMC cannot reach the perfect reliability because the estimates by PMC cannot detect the deterministic correlation unless the number of categories is equal in the variables. More studies and theoretical work in the interpretation of the estimators would enrich us.

Some typological characteristics of different estimators of the estimators described in the article are summarized in **Table 8**. Notably, again, *RPC*, *G,* and *D* are not the only options for DCERs; further studies related to such estimators as r-bireg- and r-polyreg correlations, G_2, D_2, as well as attenuation-corrected *Rit* and *eta*, as examples, would be beneficial (see footnote 6).

Known Limitations of the Treatment

The empirical section offers, obviously, just examples of what kind of effect would be obtained if an estimator with smaller quantity of deflation is used as the linking factor between the latent variables and the item. Wider comparisons of different estimators would benefit us to select most suitable estimators of correlation as the linking factors for different variables, estimators of reliability and different type of datasets. Systematic simulations also in this area would enrich us.

The DCERs in the article were given just as examples—their characteristics were not studied in-depth. Specifically, the estimators based on omega and rho are, by far, theoretical options in the settings related to factor analysis and structural equation modeling because they may require new procedures where the *outcome* of factor loadings would be (essentially) *RPC* or *G*

instead of (essentially) PMC. Notably, the current procedures of using *RPC* in EFA and SEM may *start* by using *RPC* in forming the correlation matrix, but the outcome of the loadings seems to be still, essentially, PMC. Also, Chalmers (2017) critique against the use of *RPC* in estimating reliability is worth noting. More studies in this regard would benefit us.

The study did not tackle the question of possible overestimation of reliability when using deflation-corrected estimators of reliability. *Assuming* that *RPC* does not overestimate the true correlation, it may be relevant to conclude that a deflation-corrected estimator based on *RPC* such as Eqs. (21), (24), (27), and (30) would not overestimate reliability. What would be the mechanism for overestimation? It may be possible that the estimators based on rho overestimate the reliability in the real-world settings; this would be a reasonable consequence of the results by Aquirre-Urreta et al. (2019) that rho may overestimate the true reliability with finite samples familiar in real-world testing settings with small or smallish number of test takers. From this viewpoint, the estimators based on alpha, theta and omega seem to give more conservative estimates. Theoretical and empirical studies in the area would be beneficial.

Finally, in several places in the article a loose wording concerning the deflation in the estimates of reliability was described as "remarkable" or "notable." Based on the behavior of PMC, it is expected that the effect of changing PMC with better behaving estimators of correlation in the estimators of reliability is "remarkable" or maybe even "dramatical" when the test is very easy or very demanding to the target group or with tests with incremental difficulty levels as are usual in the educational testing settings; PMC is severely deflated in these cases. Also, with the tests of incremental difficulty level where part of the test items may be very easy and part may be very demanding as is usual in the achievement testing, we may expect remarkable difference between the traditional estimators and deflation-corrected ones. However, when all items are of medium difficulty level, the effect may not be as notable. Wider empirical studies and simulations would enrich us in this regard.

DATA AVAILABILITY STATEMENT

The datasets presented in 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**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JM contributed alone in the article.

FUNDING

No specific funding was given nor applied for this study. However, it was prepared partly by the kind support of the employer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.748672/full#supplementary-material

REFERENCES

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 INDEX, AND THE GUTTMAN SCALE RESPONSE PATTERn. *Educ. Res. Perspect.* 9, 95–104.

Andrich, D., and Douglas, G. A. (1977). "Reliability: distinctions between item consistency and subject separation with the simple logistic model," in *Paper Presented at the Annual Meeting of the American Educational Research Association* (New York, NY)

Anselmi, P., Colledai, D., and Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Front. Psychol.* 10:2714. doi: 10.3389/fpsyg.2019.02714

Aquirre-Urreta, M., Rönkkö, M., and McIntosh, C. N. (2019). A cautionary note on the finite sample behavior of maximal reliability. *Psychol. Methods* 24, 236–252. doi: 10.1037/met0000176

Armor, D. (1973). Theta reliability and factor scaling. *Sociol. Methodol.* 5, 17–50. doi: 10.2307/270831

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *Br. J. Psychol.* 3, 296–322.

Chalmers, R. P. (2017). On misconceptions and the limited usefulness of ordinal alpha. *Educ. Psychol. Measurement* 78, 1056–1071. doi: 10.1177/0013164417727036

Chan, D. (2008). "So why ask me? are self-report data really that bad?," in *Statistical and Methodological Myths and Urban Legends*, eds C. E. Lance and R. J. Vanderberg (Milton Park: Routledge), doi: 10.4324/9780203867266

Cheng, Y., Yuan, K.-H., and Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educ. Psychol. Measurement* 72, 52–67. doi: 10.1177/0013164411407315

Cramer, D., and Howitt, D. (2004). *The Sage Dictionary of Statistics. A Practical Resource for Students*. Thousand Oaks, CA: SAGE Publications Inc.

Cronbach, L. J. (1951). Coefficient and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Dunn, T. J., Baguley, T., and Brunsden, V. (2013). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br. J. Psychol.* 105, 399–412. doi: 10.1111/bjop.12046

FINEEC (2018). *National Assessment of Learning Outcomes in Mathematics at Grade 9 in 2002 (Unpublished dataset opened for the re-analysis 18.2.2018)*. Helsinki: Finnish National Education Evaluation Centre (FINEEC).

Gadermann, A. M., Guhn, M., and Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Pract. Assess. Res. Eval.* 17, 1–13. doi: 10.7275/n560-j767

Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Am. Statist. Assoc.* 49, 732–764. doi: 10.1080/01621459.1954.10501231

Green, S. B., and Yang, Y. (2009). Commentary on coefficient alpha: a cautionary tale. *Psychometrika* 74, 121–135. doi: 10.1007/s11336-008-9098-4

Green, S. B., and Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educ. Measurement: Issues Practice* 34, 14–20. doi: 10.1111/emip.12100

Greene, V. L., and Carmines, E. G. (1980). Assessing the reliability of linear composites. *Sociol. Methodol.* 11, 160–17. doi: 10.2307/270862

Gulliksen, H. (1950). *Theory of Mental Tests*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika* 10, 255–282. doi: 10.1007/BF02288892

Heise, D., and Bohrnstedt, G. (1970). Validity, invalidity, and reliability. *Sociol. Methodol.* 2, 104–129. doi: 10.2307/270785

Jackson, P. H., and Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: algebraic lower bounds. *Psychometrika* 42, 567–578. doi: 10.1007/BF02295979

Jackson, R. W. B., and Ferguson, G. A. (1941). *Studies on the Reliability of Tests*. Toronto, ON: Department of Educational Research, University of Toronto.

Kaiser, H. F., and Caffrey, J. (1965). Alpha factor analysis. *Psychometrika* 30, 1–14. doi: 10.1007/BF02289743

Kendall, M. (1938). A new measure of rank correlation. *Biometrika* 30, 81–93. doi: 10.2307/2332226

Kendall, M. (1949). Rank and product–moment correlation. *Biometrika* 36, 177–193. doi: 10.2307/2332540

Kendall, M. G. (1948). *Rank Correlation Methods*, 1st Edn. London: Charles Griffin & Co Ltd.

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educ. Psychol. Measurement* 30, 61–70. doi: 10.1177/001316447003000105

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika* 2, 151–160. doi: 10.1007/BF02288391

Lavrakas, P. J. (2008). "Attenuation," in *Encyclopedia of Survey Methods*, ed. P. J. Lavrakas (Thousand Oaks, CA: Sage Publications Inc.), doi: 10.4135/9781412963947.n24

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika* 62, 245–249. doi: 10.1007/BF02295278

Li, H., Rosenthal, R., and Rubin, D. B. (1996). Reliability of measurement in psychology: from spearman-brown to maximal reliability. *Psychol. Methods* 1, 98–107. doi: 10.1037/1082-989X.1.1.98

Livingston, S. A., and Dorans, N. J. (2004). *A Graphical Approach to Item Analysis*. Research Report No. RR-04-10. Princeton, NJ: Educational Testing Service, doi: 10.1002/j.2333-8504.2004.tb01937.x

Lord, F. M. (1958). Some relations between Guttman's principal component scale analysis and other psychometric theory. *Psychometrika* 23, 291–296. doi: 10.1002/j.2333-8504.1957.tb00073.x

Lord, F. M., Novick, M. R., and Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores*. Boston, MA: Addison–Wesley Publishing Company.

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: a test of validity. *J. Market. Res.* 10, 316–318. doi: 10.2307/3149702

Martin, W. S. (1978). Effects of scaling on the correlation coefficient: additional considerations. *J. Market. Res.* 15, 304–308. doi: 10.1177/002224377801500219

McDonald, R. P. (1970). Theoretical canonical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *Br. J. Mathemat. Statist. Psychol.* 23, 1–21. doi: 10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychol. Methods* 23, 412–433. doi: 10.1037/met0000144

Meade, A. W. (2010). "Restriction of range," in *Encyclopedia of Research Design*, ed. N. J. Salkind (Thousand Oaks, CA: SAGE Publications, Inc.). doi: 10.4135/9781412961288.n309

Metsämuuronen, J. (2009). *Methods Assisting the Assessment. [Metodit arvioinnin apuna] Series Assessment of Learning Outcomes (Oppimistulosten arviointi) 1/2009*. Helsinki: Finnish National Board of Education.

Metsämuuronen, J. (2016). Item–total correlation as the cause for the underestimation of the alpha estimate for the reliability of the scale. *GJRA - Global J. Res. Anal.* 5, 471–477.

Metsämuuronen, J. (2017). *Essentials of Research Methods in Human Sciences*. Thousand Oaks, CA: SAGE Publications, Inc.

Metsämuuronen, J. (2020b). Somers' D as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. *Int. J. Educ. Methodol.* 6, 207–221. doi: 10.12973/ijem.6.1.207

Metsämuuronen, J. (2020a). Dimension-corrected Somers' D for the item analysis settings. *Int. J. Educ. Methodol.* 6, 297–317. doi: 10.12973/ijem.6.2.297

Metsämuuronen, J. (2021b). Goodman–Kruskal gamma and dimension-corrected gamma in educational measurement settings. *Int. J. Educ. Methodol.* 7, 95–118. doi: 10.12973/ijem.7.1.95

Metsämuuronen, J. (2021c). Mechanical attenuation in eta squared and some related consequences. attenuation-corrected eta and eta squared, negative values of eta, and their relation to Pearson correlation. *bioRixv [Prperint]*. doi: 10.13140/RG.2.2.29569.58723

Metsämuuronen, J. (2021d). The effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability. seeking the best options of correlation for deflation-corrected reliability. *bioRixv [Prperint]*. doi: 10.13140/RG.2.2.36496.53767/1

Metsämuuronen, J. (2021a). Directional nature of Goodman-Kruskal gamma and some consequences. identity of Goodman-Kruskal gamma and Somers delta, and their connection to Jonckheere-Terpstra test statistic. *Behaviormetrika* 48, 283–307. doi: 10.1007/s41237-021-00138-8

Metsämuuronen, J., and Ukkola, A. (2019). *Methodological Solutions of Zero Level Assessment (Alkumittauksen menetelmällisiä ratkaisuja). Publications 18:2019*. Helsinki: Finnish National Education Evaluation Centre (FINEEC).

Milanzi, E., Molenberghs, G., Alonso, A., Verbeke, G., and De Boeck, P. (2015). Reliability measures in item response theory: manifest versus latent correlation functions. *Br. J. Mathemat. Statist. Psychol.* 68, 43–64. doi: 10.1111/bmsp.12033

Moses, T. (2017). "A review of developments and applications in item analysis," in *Advancing Human Assessment. The Methodological, Psychological and Policy Contributions of ETS. Educational Testing Service*, eds R. Bennett and M. von Davier (Berlin: Springer Open), doi: 10.1007/978-3-319-58689-2_2

Novick, M. R., and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika* 32, 1–13. doi: 10.1007/BF02289400

Olsson, U. (1980). Measuring correlation in ordered two-way contingency tables. *J. Market. Res.* 17, 391–394. doi: 10.1177/002224378001700315

Pearson, K. (1896). VII. mathematical contributions to the theory of evolution. III. regression, heredity and panmixia. *Philos. Trans. R. Soc. London* 187, 253–318. doi: 10.1098/rsta.1896.0007

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution. VII. on the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. Mathematical Phys. Eng. Sci.* 195, 1–47. doi: 10.1098/rsta.1900.0022

Pearson, K. (1903). I. mathematical contributions to the theory of evolution. — XI. on the influence of natural selection on the variability and correlation of organs. *Philos. Trans. R. Soc. Mathemat. Phys. Eng. Sci.* 200, 1–66. doi: 10.1098/rsta.1903.0001

Pearson, K. (1909). On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika* 7, 96–105. doi: 10.1093/biomet/7.1-2.96

Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika* 9, 116–139. doi: 10.1093/biomet/9.1-2.116

Raykov, T. (1997a). Estimation of composite reliability for congeneric measures. *Appl. Psychol. Measurement* 21, 173–184. doi: 10.1177/01466216970212006

Raykov, T. (1997b). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau–equivalence for fixed congeneric components. *Multivariate Behav. Res.* 32, 329–354. doi: 10.1207/s15327906mbr3204_2

Raykov, T. (2004). Estimation of maximal reliability: a note on a covariance structure modeling approach. *Br. J. Mathemat. Statist. Psychol.* 57, 21–27. doi: 10.1348/000711004849295

Raykov, T. (2012). "Scale development using structural equation modeling," in *Handbook of Structural Equation Modeling*, ed. R. Hoyle (New York, NY: Guilford Press), 472–492.

Raykov, T., and Marcoulides, G. A. (2017). Thanks coefficient alpha, we still need you! *Educ. Psychol. Measurement* 79, 200–210. doi: 10.1177/0013164417725127

Revelle, W., and Condon, D. M. (2018). "Reliability," in *The Wiley Handbook of Psychometric Testing: a Multidisciplinary Reference on Survey, Scale and Test Development*, eds P. Irwing, T. Booth, and D. J. Hughes (London: John Wily & Sons).

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educ. Rev.* 9, 99–103.

Sackett, P. R., and Yang, H. (2000). Correction for range restriction: an expanded typology. *J. Appl. Psychol.* 85, 112–118. doi: 10.1037/0021-9010.85.1.112

Sackett, P. R., Lievens, F., Berry, C. M., and Landers, R. N. (2007). A cautionary note on the effect of range restriction on predictor intercorrelations. *J. Appl. Psychol.* 92, 538–544. doi: 10.1037/0021-9010.92.2.538

Schmidt, F. L., and Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence* 27, 183–198.

Schmidt, F. L., and Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, 3rd Edn. Newbury Park, CA: SAGE Publications. doi: 10.4135/9781483398105

Smith, J. K. (2005). Reconsidering reliability in classroom assessment and grading. *Educ. Measurement: Issues Practice* 22, 26–33. doi: 10.1111/j.1745-3992.2003.tb00141.x

Somers, R. H. (1962). A new asymmetric measure of correlation for ordinal variables. *Am. Sociol. Rev.* 27, 799–811. doi: 10.2307/2090408

Spearman, C. (1904). The proof and measurement of correlation between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159

Spearman, C. (1910). Correlation computed with faulty data. *Br. J. Psychol.* 3, 271–295.

Trizano-Hermosilla, I., and Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Front. Psychol.* 7:769. doi: 10.3389/fpsyg.2016.00769

Woodhouse, B., and Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: a search procedure to locate the greatest lower bound. *Psychometrika* 42, 579–591. doi: 10.1007/BF02295980

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. San Diego, CA: Mesa Press.

Yang, H. (2010). "Factor loadings," in *Encyclopedia of Research Design*, ed. N. J. Salkind (Thousand Oaks, CA: SAGE Publications), 480–483.

Yang, Y., and Green, S. B. (2011). Coefficient alpha: a reliability coefficient for the 21st century? *J. Psychoeduc. Assess.* 29, 377–392. doi: 10.1177/0734282911406668

Zumbo, B. D., Gadermann, A. M., and Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *J. Modern Appl. Statist. Methods* 6, 21–29. doi: 10.22237/jmasm/1177992180

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: 12 January 2022
doi: 10.3389/fpsyg.2021.743362

An Extension of Testlet-Based Equating to the Polytomous Testlet Response Theory Model

*Feifei Huang[1], Zhe Li[1], Ying Liu[2], Jingan Su[1], Li Yin[1] and Minqiang Zhang[1]**

[1] School of Psychology, South China Normal University, Guangzhou, China, [2] College of Teacher's Education, Guangdong University of Education, Guangzhou, China

Educational assessments tests are often constructed using testlets because of the flexibility to test various aspects of the cognitive activities and broad content sampling. However, the violation of the local item independence assumption is inevitable when tests are built using testlet items. In this study, simulations are conducted to evaluate the performance of item response theory models and testlet response theory models for both the dichotomous and polytomous items in the context of equating tests composed of testlets. We also examine the impact of testlet effect, length of testlet items, and sample size on estimating item and person parameters. The results show that more accurate performance of testlet response theory models over item response theory models was consistently observed across the studies, which supports the benefits of using the testlet response theory models in equating for tests composed of testlets. Further, results of the study indicate that when sample size is large, item response theory models performed similarly to testlet response theory models across all studies.

Keywords: testlet, test equating, item response theory model, dichotomous testlet response theory model, polytomous testlet response theory model

OPEN ACCESS

Edited by:
Marta Martín-Carbonell, Universidad Cooperativa de Colombia, Colombia

Reviewed by:
Mark D. Reckase, Michigan State University, United States
Seock-Ho Kim, University of Georgia, United States
Thomas Eckes, Ruhr University Bochum, Germany

***Correspondence:**
Minqiang Zhang
Zhangmq1117@qq.com

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

Received: *18 July 2021*
Accepted: *15 December 2021*
Published: *12 January 2022*

Citation:
Huang F, Li Z, Liu Y, Su J, Yin L and Zhang M (2022) An Extension of Testlet-Based Equating to the Polytomous Testlet Response Theory Model. Front. Psychol. 12:743362. doi: 10.3389/fpsyg.2021.743362

INTRODUCTION

In the current practice of educational measurement, test equating is a vital step to put scores from different forms onto a same scale. However, in most large-scale testing programs, it is common for a standardized test to consist of testlets (Bradlow et al., 1999; Rijmen, 2009; Cao et al., 2014; Tao and Cao, 2016). A testlet is defined as an aggregation of items which are based on a common stimulus (Wainer and Kiely, 1987; Bradlow et al., 1999). Responses to items within a testlet often tend to violate the local item independence. For example, some examinees that are more familiar with the background information covered by the testlet may have a higher probability to correctly answer the items of a specific testlet (Rijmen, 2009; Cao et al., 2014; Tao and Cao, 2016). Although researchers have conducted an abundance of studies to propose different approaches to handle local item dependence (LID), little research in the literature has focused on the performance of different approaches to accommodate LID on testlet-based test equating.

Studies have shown that the accuracy of parameter estimation produced by the testlet response theory (TRT) model is higher than the traditional item response theory (IRT) model where LID was present (Bradlow et al., 1999; Wainer and Wang, 2000; Wainer et al., 2000; Zhang, 2010; Koziol, 2016). However, numbers of studies were based on dichotomous items (Wainer and Wang, 2000; Rijmen, 2009; Cao et al., 2014). Researchers have found that although the polytomous IRT models suffer the problem of losing response pattern information, they are still much easier in interpretation and implementation (Sireci et al., 1991; Zenisky et al., 2002; Cao et al., 2014). Moreover, studies have also documented that the dichotomous IRT models could lead to misestimation of test reliability and item parameters (Sireci et al., 1991; Lawrence, 1995; Zenisky et al., 2002; Keller et al., 2003; Cao et al., 2014). Because there is little evidence about the application of TRT models for the polytomous items composed of testlets in the context of equating tests, it is not clear how the performance of TRT models might be.

It is needed to place the IRT estimates from different test forms on a common scale when conducting test equating (Kolen and Brennan, 2014). Generally, there are two kinds of parameter linking methods known as separate calibration and concurrent calibration (von Davier and von Davier, 2011; Kolen and Brennan, 2014; González and Wiberg, 2017). Separate calibration needs an equating transformation to perform the equating, while concurrent calibration can link parameters obtained from different test forms on a common scale during the estimation routine (Kolen and Brennan, 2014; González and Wiberg, 2017). Researches of test equating on the testlets were mostly based on the method of separate calibration (Lee et al., 2001; Cao et al., 2014; Tao and Cao, 2016). However, studies have shown that the accuracy of equating results of concurrent calibration were higher (Wingersky et al., 1987; Hanson and Beguin, 2002). Further, the method of concurrent calibration is easy in implementation.

Studies have investigated the influence of testlet effect size on test reliability and parameter estimates produced by IRT and TRT models (Bradlow et al., 1999; Wang et al., 2002; Zhang, 2010; Cao et al., 2014; Koziol, 2016). In addition to the testlet effect size, few studies have simultaneously investigated the impact of length of testlet items and sample size on the testlet models. However, sample size and the length of testlet items are also the important factors which can affect the accuracy of parameter estimation and equating results (Tao and Cao, 2016). Therefore, it is vital to consider those factors to compare the performance of IRT models and TRT models.

TESTLET RESPONSE THEORY MODELS

Bradlow et al. (1999) first proposed a dichotomous testlet item response model, which is based on the two-parameter logistic model (2PLM) and incorporated the random item testlet effect parameter. Since then, the TRT models have been introduced in a series of papers (Wainer et al., 2000, 2007; Wainer and Wang, 2001; Wang et al., 2002, 2004; Wang and Wilson, 2005). Researchers have found that the TRT models are predominantly used to represent the multidimensional IRT approach to model LID due to testlet effects (DeMars, 2006; Li et al., 2006).

The 2PLM and the two-parameter TRT model (2PTM) can be expressed as:

$$\mathrm{P}\left(y_{ij} = 1\right) = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} \tag{1}$$

$$\mathrm{P}\left(y_{ij} = 1\right) = \frac{\exp[a_j(\theta_i - b_j - \gamma_{id(j)})]}{1 + \exp[a_j(\theta_i - b_j - \gamma_{id(j)})]}, \tag{2}$$

where a_j is the discrimination parameter for item j, b_j is the difficulty parameter for item j, θ_i is the latent trait level for examinee i. For the TRT model, $d(j)$ denotes a testlet containing item j, $\gamma_{id(j)}$ is the random effect for examinee i on testlet $d(j)$, which describes the interaction between examinee's performance on the testlet and items (LID) within the testlet. The model assumes that$\gamma_{id(j)} \sim N[0, \sigma^2_{\gamma_{id(j)}}]$. Note that$\sigma^2_{\gamma_{id(j)}}$ reflects the amount of the testlet effect. The larger the $\sigma^2_{\gamma_{id(j)}}$ is, the larger the testlet effect will be.

However, the increasing number of educational tests consisted of polytomous item have received a substantial amount of attention because of the need for more realistic and richer forms of assessment. Therefore, researchers extended the graded response model (GRM) which is widely used to a graded response testlet model (Wang et al., 2002). Further, they developed the corresponding software SCORIGHT3.0 to estimate parameters by using the Monte Carlo method within the Bayesian framework (Wang et al., 2004).

The GRM and graded response testlet model (GRTM) can be expressed as:

$$P^*_{mx}(\theta) = \frac{\exp\left[\alpha_m\left(\theta_i - b_{mx}\right)\right]}{1 + \exp\left[\alpha_m\left(\theta_i - b_{mx}\right)\right]} (x = 0, 1, 2, \ldots, k_m - 1) \tag{3}$$

$$P_{mx}(\theta) = P^*_{mx}(\theta) - P^*_{m(x+1)}(\theta) \tag{4}$$

$$P^*_{jn}(\theta) = \frac{\exp\left[\alpha_j\left(\theta_i - b_{jn} - \gamma_{id(j)}\right)\right]}{1 + \exp\left[\alpha_j\left(\theta_i - b_{jn} - \gamma_{id(j)}\right)\right]} (n = 0, 1, 2, \ldots, k_j - 1) \tag{5}$$

$$P_{jn}(\theta) = P^*_{jn}(\theta) - P^*_{j(n+1)}(\theta), \tag{6}$$

where $P^*_{mx}(\theta)$ is the probability of an examinee with a given θ responding to category x or higher of item m, a_m is the discrimination parameter for item m, b_{mx} is the category boundary for score x on item m, $P_{mx}(\theta)$ is the probability of an examinee with a given θ will score in a particular category of item m. Compared with the GRM, $d(j)$ denotes a testlet containing item j, $\gamma_{id(j)}$ is the random effect for examinee i on testlet $d(j)$.

The Present Study

This paper presents the results of two simulation studies that addresses these two issues. First, the performance of IRT

models and TRT models for the dichotomous and polytomous items by using the concurrent calibration in the context of equating tests composed of testlets was assessed. Second, the effect of testlet effect, sample size and length of testlet items on parameter estimates produced by IRT and TRT models was investigated.

The rest of this article is organized as follows. First, the IRT and TRT models for the dichotomous and polytomous items are briefly introduced. Second, two simulation studies are conducted to assess the IRT models and TRT models. These simulations also demonstrate how testlet effect, length of testlet items, and sample size affect item and person parameters estimation. Finally, this article draws conclusions for the performance of IRT models and TRT models and suggestions for future study are provided.

MATERIALS AND METHODS

Study Design

The simulation study employed the non-equivalent anchor test (NEAT) design. In the NEAT design, two simulations with several manipulated factors were conducted to compare the performance of item response models and testlet response models in the context of equating tests composed of testlets (as shown in **Table 1**). For the first simulation study, four major independent variables were manipulated: (a) models (2PLM and 2PTM), (b) testlet effect (0.5, 1, and 2), (c) length of testlet items (5 and 10), and (d) sample size (1,000 and 2,000 examinees). The manipulated factors for the second simulation study were as same as the first one, except for the models. As the purpose of the second simulation study was to compare the performance of polytomous item response models and testlet response models, the GRM and the GRTM were selected.

Simulation Process

Six pairs of test forms were created with varying degree of testlet effect and different length of testlet items for each simulation research. Each test pair was consisted of a base form and a new form. Each test form had a total of 60 multiple choice items in the first simulation study or 60 polytomous items in the second simulation study, composed of 40 non-anchor items and 20 anchor items. For the non-anchor items, there were 20 locally independent items, and 4 testlets with 5 items per testlet or 2 testlets with 10 items per testlet. For the anchor items, there were 2 testlets with 5 items per testlet or 1 testlets with 10 items per testlet.

Item parameters for the base form and the new form composed solely of locally independent non-anchor items were randomly selected from the same population distributions. Specifically, $\ln a \sim N(0, 1)$, constrained to (0, 2.5); $b \sim N(0, 1)$, constrained to (-3, +3). Item parameters of independent anchor items which were shared by the base form and the new form were also randomly selected from the same population distributions. Specifically, $\ln a \sim N(0, 1)$, constrained to (1, 2); $b \sim N(0, 1)$, constrained to (-3, +3). The population distribution to the *a*-parameter for the anchor items with a slightly higher than the non-anchor items is to assure the representative of the anchor items (Wang et al., 2002; Cao et al., 2014; Tao and Cao, 2016). The ability population distribution for the base form is the same as the *b*-parameter population distribution to assure that the difficulty of test is appropriate for the examinees (Tao and Cao, 2016). The ability population distribution for the new form with a slightly higher mean than the base form to reflect ability differences between two groups (Lee et al., 2016; Andersson, 2018). For the three pairs of test forms, the testlet effect indexed by the $\sigma^2_{\gamma_{id(j)}}$ for the base form and the reference form were drawn from three uniform distributions: (0.1, 0.5), (0.6, 1), and (1.1, 2.0) corresponding to low, moderate and high levels of LID, respectively (Wang et al., 2002; DeMars, 2006; Cao et al., 2014; Tao and Cao, 2016).

The probability of each examinee's response to each item was calculated based on the simulated parameters mentioned above using the 2PL model, the 2PTM, the GRM, and the GRTM, respectively. Then, the probability was compared with a number randomly drawn from U (0, 1). For the dichotomous items, if the probability was larger than the random number, the response was coded as "1"; otherwise, as "0." For the polytomous items, if the random number was larger than the

TABLE 1 | Summary of the study design for the two simulation studies.

	Manipulated factors			
The first simulation study	Models	2PLM	2PTM	
	Testlet effect	0.5	1	2
	Length of testlet items	5	10	
	Sample size	1,000	2,000	
The second simulation study	Models	GRM	GRTM	
	Testlet effect	0.5	1	2
	Length of testlet items	5	10	
	Sample size	1,000	2,000	

TABLE 2 | Statistical summary of the discrimination parameter for the dichotomous testlet items.

Evaluation criteria	Model	Sample size	Length of testlet = 5			Length of testlet = 10		
			Testlet effect					
			0.5	1	2	0.5	1	2
RMSE	2PLM	1,000	0.51	0.73	1.22	0.56	0.80	1.24
		2,000	0.46	0.63	0.92	0.50	0.76	1.12
	2PTM	1,000	0.35	0.43	0.46	0.36	0.45	0.49
		2,000	0.32	0.40	0.42	0.29	0.43	0.47
Bias	2PLM	1,000	0.42	0.58	0.97	0.46	0.72	1.07
		2,000	0.39	0.53	0.91	0.41	0.70	1.02
	2PTM	1,000	0.20	0.25	0.31	0.21	0.30	0.27
		2,000	0.19	0.23	0.28	0.18	0.25	0.28
SEE	2PLM	1,000	0.29	0.44	0.74	0.32	0.35	0.63
		2,000	0.24	0.34	0.14	0.29	0.30	0.46
	2PTM	1,000	0.29	0.35	0.34	0.29	0.34	0.41
		2,000	0.26	0.33	0.31	0.23	0.35	0.38

TABLE 3 | Statistical summary of the difficulty parameter for the dichotomous testlet items.

Evaluation criteria	Model	Sample size	Length of testlet = 5			Length of testlet = 10		
			Testlet effect					
			0.5	1	2	0.5	1	2
RMSE	2PLM	1,000	0.22	0.23	0.29	0.23	0.27	0.34
		2,000	0.21	0.22	0.28	0.20	0.23	0.30
	2PTM	1,000	0.21	0.13	0.16	0.14	0.14	0.17
		2,000	0.12	0.11	0.13	0.12	0.13	0.16
Bias	2PLM	1,000	−0.10	−0.12	−0.15	−0.11	−0.14	−0.18
		2,000	−0.09	−0.10	−0.13	−0.10	−0.13	−0.16
	2PTM	1,000	−0.10	−0.08	−0.07	−0.08	−0.08	−0.07
		2,000	−0.07	−0.07	−0.06	−0.07	−0.06	−0.06
SEE	2PLM	1,000	0.20	0.20	0.25	0.20	0.23	0.29
		2,000	0.19	0.20	0.25	0.17	0.19	0.25
	2PTM	1,000	0.20	0.20	0.25	0.20	0.23	0.29
		2,000	0.10	0.08	0.12	0.10	0.12	0.15

TABLE 4 | Statistical summary of the ability parameter for the dichotomous testlet items.

Evaluation criteria	Model	Sample size	Length of testlet = 5			Length of testlet = 10		
			Testlet effect					
			0.5	1	2	0.5	1	2
RMSE	2PLM	1,000	0.24	0.27	0.31	0.26	0.31	0.37
		2,000	0.22	0.26	0.27	0.24	0.29	0.34
	2PTM	1,000	0.18	0.21	0.23	0.20	0.21	0.27
		2,000	0.17	0.19	0.22	0.18	0.20	0.24
Bias	2PLM	1,000	−0.11	−0.12	0.14	−0.13	−0.15	−0.19
		2,000	−0.10	−0.11	0.12	−0.11	−0.14	−0.15
	2PTM	1,000	−0.08	−0.08	0.07	−0.07	−0.08	−0.07
		2,000	−0.07	−0.06	0.06	−0.06	−0.07	−0.06
SEE	2PLM	1,000	0.21	0.24	0.28	0.23	0.27	0.32
		2,000	0.20	0.24	0.24	0.21	0.25	0.31
	2PTM	1,000	0.16	0.19	0.22	0.19	0.19	0.26
		2,000	0.15	0.18	0.21	0.17	0.19	0.23

TABLE 5 | Statistical summary of the discrimination parameter for the polytomous testlet items.

Evaluation criteria	Model	Sample size	Length of testlet = 5			Length of testlet = 10		
			Testlet effect					
			0.5	1	2	0.5	1	2
RMSE	GRM	1,000	0.65	0.69	0.97	0.67	0.72	1.04
		2,000	0.61	0.63	0.91	0.67	0.70	0.95
	GRTM	1,000	0.38	0.37	0.32	0.32	0.33	0.29
		2,000	0.36	0.34	0.31	0.31	0.32	0.29
Bias	GRM	1,000	0.69	0.74	1.03	0.73	0.83	1.12
		2,000	0.68	0.73	1.02	0.71	0.79	1.04
	GRTM	1,000	0.41	0.43	0.39	0.34	0.36	0.45
		2,000	0.38	0.37	0.39	0.32	0.35	0.42
SEE	GRM	1,000	0.23	0.27	0.35	0.29	0.41	0.42
		2,000	0.30	0.37	0.46	0.23	0.37	0.42
	GRTM	1,000	0.15	0.22	0.22	0.11	0.14	0.34
		2,000	0.12	0.15	0.24	0.08	0.14	0.30

cumulative probability with category 1, the response was coded as "0"; if the random number was between the cumulative probability with category 1 and the cumulative probability with category 2, the response was coded as "1"; if the random number was between the cumulative probability with category 2 and the cumulative probability with category 3, the response was coded as "2"; if the random number was between the cumulative probability with category 3 and the cumulative probability with category 4, the response was coded as "3"; otherwise, as "4."

TABLE 6 | Statistical summary of the difficulty parameter for the polytomous testlet items.

Evaluation criteria	Model	Sample size	Length of testlet	Testlet effect											
				0.5				1				2			
				b_1	b_2	b_3	b_4	b_1	b_2	b_3	b_4	b_1	b_2	b_3	b_4
RMSE	GRM	1,000	5	0.35	0.20	0.24	0.43	0.36	0.21	0.25	0.45	0.41	0.23	0.28	0.49
			10	3.49	1.89	1.74	3.22	3.21	1.75	1.61	2.98	2.99	1.61	1.49	2.76
		2,000	5	0.33	0.18	0.22	0.42	0.35	0.20	0.23	0.44	0.40	0.22	0.26	0.48
			10	3.45	1.86	1.76	3.24	3.22	1.74	1.62	2.99	2.96	1.61	1.49	2.73
	GRTM	1,000	5	0.27	0.19	0.15	0.16	0.33	0.18	0.24	0.26	0.22	0.16	0.16	0.20
			10	0.35	0.20	0.25	0.44	0.36	0.21	0.25	0.45	0.40	0.23	0.27	0.49
		2,000	5	0.27	0.17	0.14	0.12	0.18	0.14	0.12	0.12	0.19	0.14	0.13	0.16
			10	0.35	0.18	0.24	0.41	0.37	0.20	0.24	0.44	0.41	0.23	0.25	0.47
Bias	GRM	1,000	5	0.22	0.02	−0.15	−0.35	0.23	0.02	0.15	−0.37	0.30	0.04	−0.16	−0.41
			10	2.42	0.94	−0.63	−2.03	2.66	0.86	0.58	−2.18	2.89	0.78	−0.54	−2.39
		2,000	5	0.20	0.01	−0.14	−0.29	0.21	0.01	0.13	−0.36	0.27	0.02	−0.14	−0.38
			10	2.40	0.92	−0.67	−2.05	2.64	0.87	−0.60	−2.16	2.84	0.76	−0.54	−2.36
	GRTM	1,000	5	−0.17	−0.11	−0.07	−0.01	−0.13	−0.10	−0.07	−0.05	−0.08	−0.09	−0.10	−0.11
			10	0.23	0.02	−0.15	−0.36	0.24	0.02	−0.16	−0.36	0.29	0.04	−0.19	−0.43
		2,000	5	−0.15	−0.09	−0.07	−0.01	−0.14	−0.11	−0.09	−0.05	−0.07	−0.09	−0.09	−0.10
			10	0.22	0.02	−0.14	−0.33	0.23	0.03	−0.15	−0.36	0.27	0.05	−0.16	−0.40
SEE	GRM	1,000	5	0.27	0.20	0.19	0.25	0.28	0.21	0.20	0.26	0.28	0.23	0.23	0.27
			10	2.51	1.64	1.62	2.50	1.80	1.52	1.50	2.03	0.77	1.41	1.39	1.38
		2,000	5	0.19	0.06	0.17	0.40	0.30	0.14	0.20	0.42	0.35	0.17	0.23	0.45
			10	2.48	1.62	1.63	2.51	1.84	1.51	1.50	2.07	0.83	1.42	1.39	1.37
	GRTM	1,000	5	0.21	0.15	0.13	0.16	0.30	0.15	0.23	0.26	0.20	0.13	0.12	0.17
			10	0.26	0.20	0.20	0.25	0.27	0.21	0.19	0.27	0.28	0.23	0.19	0.23
		2,000	5	0.22	0.14	0.12	0.12	0.11	0.09	0.08	0.11	0.18	0.11	0.09	0.12
			10	0.27	0.18	0.19	0.24	0.29	0.20	0.19	0.25	0.31	0.22	0.19	0.25

TABLE 7 | Statistical summary of the ability parameter for the polytomous testlet items.

Evaluation criteria	Model	Sample size	Length of testlet = 5			Length of testlet = 10		
			Testlet effect					
			0.5	1	2	0.5	1	2
RMSE	GRM	1,000	0.22	0.25	0.32	0.25	0.30	0.37
		2,000	0.21	0.24	0.28	0.24	0.29	0.32
	GRTM	1,000	0.22	0.22	0.23	0.22	0.24	0.24
		2,000	0.20	0.21	0.23	0.21	0.22	0.23
Bias	GRM	1,000	−0.08	−0.11	−0.14	−0.10	−0.13	−0.19
		2,000	−0.07	−0.09	−0.13	−0.09	−0.13	−0.17
	GRTM	1,000	−0.09	−0.08	−0.08	−0.09	−0.08	−0.09
		2,000	−0.08	−0.06	−0.07	−0.07	−0.07	−0.09
SEE	GRM	1,000	0.20	0.22	0.29	0.23	0.27	0.32
		2,000	0.20	0.22	0.25	0.22	0.26	0.27
	GRTM	1,000	0.20	0.20	0.22	0.20	0.23	0.22
		2,000	0.18	0.20	0.22	0.20	0.21	0.21

TABLE 8 | Parameters of the dichotomous items for the reference form and new form.

	Reference form					New form			
	Items	**Testlets**	**a**	**b**		**Items**	**Testlets**	**a**	**b**
Non-anchor items	1		1.53	0.18	Non-anchor items	1		1.45	−0.59
	2		2.32	−1.49		2		1.04	−0.17
	3		1.04	−1.62		3		1.83	−1.19
	4		1.71	0.89		4		1.25	−1.62
	5		1.47	0.63		5		1.40	1.63
	6		1.68	−1.22		6		1.52	−1.22
	7		1.49	−0.07		7		1.42	−1.47
	8		1.48	−0.04		8		1.53	−0.51
	9		2.20	−1.05		9		1.92	1.68
	10		1.13	−2.05		10		1.64	0.55
	11		1.04	−0.56		11		1.37	−0.34
	12		1.11	−1.30		12		1.56	1.67
	13		1.28	0.78		13		1.46	0.85
	14		1.53	0.91		14		1.58	0.32
	15		1.35	0.67		15		1.12	0.17
	16		1.68	−1.58		16		1.62	−0.03
	17		1.73	0.43		17		1.94	0.20
	18		1.64	0.65		18		1.02	0.12
	19		1.11	0.15		19		1.53	0.50
	20		1.85	−0.57		20		1.23	−0.93
	21	1	1.09	1.35		21	1	1.52	1.64
	22	1	2.32	1.53		22	1	1.67	−1.11
	23	1	1.52	0.97		23	1	1.68	−1.11
	24	1	0.81	0.54		24	1	1.70	−0.71
	25	1	1.25	1.05		25	1	0.24	0.19
	26	2	1.39	0.01		26	1	1.12	−0.45
	27	2	1.91	−1.43		27	1	1.43	0.62
	28	2	1.48	1.76		28	1	0.68	−0.40
	29	2	2.36	−0.44		29	1	1.68	0.72
	30	2	1.76	−0.61		30	1	2.34	−1.86
	31	3	1.04	−1.05		31	2	1.74	−0.83
	32	3	2.15	−0.86		32	2	1.70	0.65
	33	3	1.75	0.24		33	2	0.80	0.81
	34	3	1.35	−0.55		34	2	1.75	0.27
	35	3	1.78	0.66		35	2	1.62	−0.19
	36	4	1.91	−0.99		36	2	1.72	0.87
	37	4	1.63	2.87		37	2	0.31	−1.49
	38	4	1.12	−0.07		38	2	0.96	−0.38
	39	4	1.23	−0.68		39	2	1.24	0.02
	40	4	0.75	0.11		40	2	1.78	0.17
Anchor items	41	5	1.45	−0.59	Anchor items	41	3	1.46	−0.15
	42	5	1.04	−0.17		42	3	1.51	−0.64
	43	5	1.83	−1.19		43	3	1.87	1.10
	44	5	1.25	−1.62		44	3	1.67	0.27
	45	5	1.40	1.63		45	3	1.53	−0.43
	46	6	1.52	−1.22		46	3	1.24	0.40
	47	6	1.42	−1.47		47	3	1.52	−1.53
	48	6	1.53	−0.51		48	3	1.52	−0.52
	49	6	1.92	1.68		49	3	1.69	0.01
	50	6	1.64	0.55		50	3	1.82	1.31
	51		1.37	−0.34		51		1.56	−1.62

(Continued)

TABLE 8 | (Continued)

	Reference form				New form			
	Items	**Testlets**	**a**	**b**	**Items**	**Testlets**	**a**	**b**
	52		1.56	0.61	52		1.46	−0.06
	53		1.46	0.85	53		1.47	−0.35
	54		1.58	0.32	54		1.61	0.86
	55		1.12	0.17	55		1.42	−0.73
	56		1.62	−0.03	56		1.78	−0.84
	57		1.94	0.20	57		2.00	−0.83
	58		1.02	0.12	58		2.14	0.42
	59		1.53	0.50	59		1.56	−0.39
	60		1.23	−0.93	60		1.80	−1.70
Mean			1.51	−0.06			1.50	−0.12
Standard deviation			0.36	1.00			0.38	0.92

An R (version 3.3.1, R Core Team, 2016) program was written to generate data and calibrate the response data by the 2PL model, the 2PTM, the GRM, and the GRTM, respectively. The program flexMIRT (Cai, 2017) were used to conduct the concurrent calibration. Related R codes could be requested from the correspondence author.

Evaluation Criteria

The focus of our study was not only on comparing IRT and TRT models, but also on the effect of testlet effect, sample size and length of testlet items on parameter estimates produced by IRT and TRT models. Therefore, we used the equating bias, standard error of equating and root mean square error to assess the performance of IRT models and TRT models for the dichotomous and polytomous items in the context of equating tests composed of testlets. Bias is an indicator of systematic error in equating. SEE is an indicator of random sampling error in equating. RMSE represents the total error in equating, which were defined as

$$Bias = \frac{1}{R}\sum_{r=1}^{R}\widehat{\lambda}-\lambda \tag{7}$$

$$SEE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\lambda}-\overline{\widehat{\lambda}}\right)^2} \tag{8}$$

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\lambda}-\lambda\right)^2}, \tag{9}$$

where $\widehat{\lambda}$ and λ were the estimated and true values for item parameters and ability parameter, R was the total number of replications (Each condition was replicated 500 times in this study), and $\overline{\widehat{\lambda}}$ was the average of $\widehat{\lambda}$ over the R replications.

RESULTS

Tables 2, **3** summarize the results of computing the RMSE, bias and SEE of equating accuracy of the discrimination parameter and difficulty parameter for the dichotomous testlet item. In terms of the bias and SEE, it is clear that the values of 2PTM were smaller than that of 2PLM. The discrimination parameters were overestimated for all conditions, but the difficulty parameters were underestimated. With regard to the RMSE, the RMSE values of 2PTM were smaller than that of 2PLM across all simulation conditions. Besides, a large sample size resulted in a smaller bias and RMSE. The bias, SEE and RMSE of 2PLM increased as the testlet effect and the length of testlet increased. However, no systematic patterns were observed for the bias, SEE and RMSE of 2PTM as the testlet effect and the length of testlet increased. In summary, the 2PTM had higher equating accuracy than the 2PLM for the discrimination parameter and difficulty parameter under different simulation conditions. Further, both two models could reduce the equating error with a larger sample.

The values of RMSE, bias and SEE of the ability parameter for the dichotomous testlet item across all simulation conditions are presented in **Table 4**. In terms of the bias, the ability parameter was underestimated under different conditions. A short length of testlet and a small testlet effect were associated with a more precise estimation of the ability parameter. In addition, similar trends can also be observed that the RMSE, bias and SEE decreased as the sample size increased. On the whole, the 2PTM performed better than the 2PLM.

Table 5 summarizes the results of computing the RMSE, bias and SEE of equating accuracy of the discrimination parameter for the polytomous testlet item. In terms of the bias and SEE, it is clear that the values of GRTM were smaller than that of GRM. The discrimination parameters were overestimated for all conditions. With regard to the RMSE, the RMSE values of GRTM were smaller than that of GRM across all simulation conditions. The same findings for the dichotomous testlet item applied to the polytomous testlet item, as evidenced by the results that a long length of testlet (i.e., 10) and a high testlet effect (i.e., 2) resulted in a larger RMSE of GRTM. In addition, the patterns of the bias, SEE and RMSE of GRTM were the same as those in the dichotomous testlet item. Additionally,

TABLE 9 | Parameters of the polytomous items for the reference form and new form.

	Reference form								New form							
	Items	**Testlets**	**a**	**b1**	**b2**	**b3**	**b4**		**Items**	**Testlets**	**a**	**b1**	**b2**	**b3**	**b4**
Non-anchor items	1		0.59	−0.03	0.27	0.50	0.56	Non-anchor items	1		1.21	−1.40	−1.17	−0.24	0.80
	2		0.59	−2.45	−1.52	−0.96	0.10		2		1.12	−2.00	−0.72	0.05	0.06
	3		1.39	−1.38	−0.17	0.47	1.73		3		1.13	−1.02	−0.61	−0.35	0.15
	4		1.20	−0.70	0.35	0.77	1.01		4		1.18	−1.07	−0.86	−0.19	0.44
	5		1.31	−0.20	0.15	0.63	2.07		5		1.27	−1.17	−0.76	−0.20	1.60
	6		0.88	−0.49	0.00	0.36	1.19		6		1.37	−0.67	−0.66	−0.35	0.18
	7		0.83	−1.01	−0.46	−0.30	0.70		7		1.21	−2.63	−0.53	1.35	2.56
	8		0.68	−1.00	0.46	0.98	1.35		8		1.21	−1.10	0.04	0.42	0.44
	9		1.17	−0.32	0.00	0.23	1.71		9		1.01	−0.57	0.32	0.50	1.23
	10		1.36	−0.93	−0.76	−0.52	1.04		10		1.46	0.13	0.90	0.96	1.01
	11		1.28	−1.68	−0.99	−0.48	1.33		11		1.11	−0.01	0.92	1.58	2.47
	12		0.74	−1.17	−0.73	−0.69	0.76		12		1.31	−0.92	−0.68	0.43	0.76
	13		1.26	−1.34	−1.15	0.22	0.52		13		1.59	−1.12	−1.09	−0.52	0.21
	14		1.00	−0.58	−0.05	0.29	1.20		14		1.74	0.22	0.37	1.14	1.14
	15		1.37	−0.36	−0.09	0.81	1.02		15		1.18	−0.15	0.54	0.86	0.92
	16		0.74	−1.86	−0.21	0.12	1.26		16		1.62	−0.53	−0.07	0.07	0.64
	17		0.80	−0.89	−0.32	0.37	0.73		17		1.20	−0.47	0.15	0.85	1.50
	18		0.74	−1.67	−0.72	0.62	1.32		18		1.22	−0.89	−0.21	0.52	0.66
	19		1.31	−0.92	−0.65	0.41	1.66		19		1.18	−0.90	−0.84	−0.75	0.75
	20		0.86	−1.04	−0.78	−0.15	1.28		20		1.46	−1.62	−0.83	0.12	1.22
	21	1	1.00	−0.59	0.16	0.20	0.64		21	1	0.98	−1.18	−0.93	−0.66	1.25
	22	1	1.60	−0.79	−0.09	0.05	1.46		22	1	1.06	−0.35	−0.19	0.37	0.87
	23	1	1.03	−1.48	−0.64	−0.46	0.46		23	1	0.35	−0.87	−0.42	0.42	0.98
	24	1	0.59	−0.40	1.22	1.23	2.06		24	1	0.79	−1.42	−0.90	−0.47	0.49
	25	1	0.89	−0.88	−0.40	1.03	2.53		25	1	1.11	−0.30	−0.16	−0.16	0.61
	26	2	1.00	−1.29	−0.94	−0.38	0.77		26	1	1.29	−1.28	−0.97	−0.85	0.74
	27	2	1.24	−2.03	−1.34	−0.05	0.59		27	1	1.32	−0.87	−0.31	0.50	1.01
	28	2	1.13	−0.88	−0.47	0.07	1.06		28	1	1.44	−1.04	0.27	0.71	1.54
	29	2	0.60	−0.97	−0.67	−0.03	0.71		29	1	0.47	−1.36	−0.81	−0.37	−0.16
	30	2	0.97	−0.62	0.01	0.94	1.59		30	1	1.11	0.13	0.41	1.10	1.32
	31	3	1.23	−1.20	−1.01	0.64	0.80		31	2	0.71	−0.88	−0.03	0.06	0.42
	32	3	1.07	−0.36	−0.34	−0.32	0.54		32	2	0.85	−0.95	0.26	2.02	2.54
	33	3	0.36	−1.04	−0.49	0.14	1.00		33	2	1.18	−1.08	−0.85	−0.07	0.56
	34	3	0.76	−1.68	0.69	0.74	1.01		34	2	1.08	−1.17	0.30	0.98	1.70
	35	3	0.90	0.45	0.72	1.16	1.69		35	2	1.15	−1.24	−0.40	0.91	1.55
	36	4	1.80	−1.81	0.14	0.96	1.19		36	2	0.46	−0.70	−0.18	0.52	0.84
	37	4	1.13	0.66	0.88	1.01	1.93		37	2	0.55	−2.01	−0.78	−0.38	−0.10
	38	4	1.26	−1.20	−0.15	0.44	0.79		38	2	1.68	−1.45	−0.31	−0.22	0.22
	39	4	1.00	−2.07	0.09	0.11	0.19		39	2	0.90	−2.05	−0.59	1.51	1.60
	40	4	0.56	−1.26	−0.38	0.23	0.62		40	2	1.98	−0.09	0.99	1.98	2.42
Anchor items	41	5	1.21	−1.40	−1.17	−0.24	0.80	Anchor items	41	3	0.93	−1.19	−1.14	−0.15	0.23
	42	5	1.12	−2.00	−0.72	0.05	0.06		42	3	1.57	−0.01	0.20	1.20	1.26
	43	5	1.13	−1.02	−0.61	−0.35	0.15		43	3	0.86	−2.01	0.25	0.75	1.07
	44	5	1.18	−1.07	−0.86	−0.19	0.44		44	3	0.69	−0.33	0.79	0.96	1.07
	45	5	1.27	−1.17	−0.76	−0.20	1.60		45	3	1.22	−0.94	−0.27	0.68	0.77
	46	6	1.37	−0.67	−0.66	−0.35	0.18		46	3	0.89	−0.94	−0.68	0.48	2.29
	47	6	1.21	−2.63	−0.53	1.35	2.56		47	3	0.80	−1.19	−1.11	−0.40	0.19
	48	6	1.21	−1.10	0.04	0.42	0.44		48	3	1.30	−1.50	−1.03	−0.21	0.71
	49	6	1.01	−0.57	0.32	0.50	1.23		49	3	1.03	−0.39	0.28	0.29	0.35
	50	6	1.46	0.13	0.90	0.96	1.01		50	3	1.11	−2.57	−1.05	−0.52	1.02
	51		1.11	−0.01	0.92	1.58	2.47		51		0.94	−1.57	0.05	0.51	0.90

(Continued)

TABLE 9 | (Continued)

	Reference form							New form						
	Items	Testlets	a	b1	b2	b3	b4	Items	Testlets	a	b1	b2	b3	b4
	52		1.31	−0.92	−0.68	0.43	0.76	52		0.64	−2.00	0.75	0.79	1.30
	53		1.59	−1.12	−1.09	−0.52	0.21	53		1.07	−0.61	−0.48	1.65	2.36
	54		1.74	0.22	0.37	1.14	1.14	54		0.67	−0.47	0.00	0.10	0.35
	55		1.18	−0.15	0.54	0.86	0.92	55		1.36	0.52	0.59	1.17	1.24
	56		1.62	−0.53	−0.07	0.07	0.64	56		0.97	−0.27	0.00	0.30	0.40
	57		1.20	−0.47	0.15	0.85	1.50	57		0.68	−2.02	−0.25	0.51	1.23
	58		1.22	−0.89	−0.21	0.52	0.66	58		0.54	−0.69	−0.23	0.03	1.01
	59		1.18	−0.90	−0.84	−0.75	0.75	59		1.01	−1.74	−1.33	−1.30	0.54
	60		1.46	−1.62	−0.83	0.12	1.22	60		0.70	−1.95	0.62	0.67	0.84
Mean			1.10	−0.96	−0.27	0.29	1.05			1.10	−1.00	−0.26	0.36	0.97
SD			0.31	0.67	0.61	0.58	0.60			0.33	0.69	0.61	0.72	0.66

a large sample size resulted in a smaller bias and RMSE for both GRM and GRTM.

Regarding the difficulty parameter for the polytomous testlet item, as shown in **Table 6**, with regard to the bias, SEE and RMSE, a long testlet length was associated with a less precise estimation of the difficulty parameter, and testlet effect and sample size had a trivial impact on the difficulty parameter for the GRTM. Additionally, the results were consistent across all categories. On the contrary, the difficulty parameter estimation was worse with a longer testlet length and a larger testlet effect for the GRM. Furthermore, the difficulty parameter estimation of the category 1 and category 4 were more deteriorated compared with the category 2 and category 3 for the GRM. Similarly, the sample size had a trivial effect on the on the difficulty parameter for the GRM. In summary, the GRTM performed better than the GRM.

For the ability parameter, as shown in **Table 7**, the parameter was underestimated under different conditions as indicated by the bias. The same findings for the dichotomous testlet item applied to the polytomous testlet item, as evidenced by the results that a short length of testlet and a small testlet effect were associated with a more precise estimation of the ability parameter. Additionally, a large sample size resulted in a smaller bias, SEE and RMSE for both GRM and GRTM.

DISCUSSION

In this study, we compared the performance of IRT models and TRT models for the dichotomous and polytomous items in the context of equating tests composed of testlets. For achieving the most generalization, in this study, the 2PL and the TRT model were selected as the item response functions for the dichotomous items, and the GRM and GRTM model were selected as the item response functions for the polytomous items. In addition, several factors were examined through the simulation studies including (a) testlet effect, (b) length of testlet items, and (c) sample size.

The simulation results showed that the TRT model always performed much more better than 2PL model when LID was present across all the test conditions. Previous studies had demonstrated that the TRT model could provide more flexibility and accuracy to the testlet-based test equating (Bradlow et al., 1999; Wainer et al., 2000; DeMars, 2006; Cao et al., 2014). Further, in addition to the confirmation of previous findings, one important contribution of this study was that a comparison was made between GRM and GRTM, which was an extension of testlet-based equating to the polytomous testlet response theory model. Despite the growing recognition of the testlet-based equating, the polytomous testlet response theory model has received little attention in the literature. Comparisons made in this study showed that the GRTM yielded more accurate item parameter estimates than the GRM when LID was present. One possible explanation could be that the GRTM, as a development from the GRM, provides more accuracy to model testlet-based tests. Therefore, use of the TRT-based models is recommended for both the dichotomous and polytomous items as they will minimize the impact of LID on the testlet-based equating.

Moreover, as in the simulation study, several factors were examined. In terms of testlet effect, it was seen that both the 2PL model and GRM were more sensitive, whereas the TRT model and GRTM seemed relatively robust as testlet effect increased from low to high. This general pattern has been consistently observed in the previous study with the comparison of different IRT models on testlet-based test equating for the dichotomous items (Cao et al., 2014), but the previous study has not taken the polytomous items into consideration. Concerning the length of testlet items, it was clear, as discussed earlier, that the TRT model and GRTM were more accurate as the length of testlet items increased than were the 2PL model and GRM. More specifically, the 2PL model and GRM consistently revealed a substantial amount of bias in parameter estimating, which led to larger overall equating errors. This may be the case because both the 2PL model and GRM could lead to the misestimation of item parameters when they were used to handle the LID caused by testlet (Zenisky et al., 2002; Keller et al., 2003). Under the NEAT design, both IRT-based models and TRT-based models tended to have smaller errors with a larger sample size primarily due to the

reduced errors of parameter estimating. Given the fact that most equating procedures require large samples for accurate estimates (Kolen and Brennan, 2014; Babcock and Hodge, 2019).

Although the current research successfully used the concurrent calibration to compare the performance of IRT models and TRT models for the dichotomous and polytomous items in the context of equating tests composed of testlets, it is not without limitations. First, there are various polytomous IRT models, such as the nominal response model, or the generalized partial credit model (Nering and Ostini, 2010; van der Linden, 2016). More research is needed to compare polytomous items with other models in the context of equating tests composed of testlets. Second, this article considered two particular test formats: dichotomous items and polytomous items, respectively. In practice, the test format (e.g., mixed-format tests) might be more complex depending on the purpose of the test (von Davier and Wilson, 2007). Future research could focus on testlet-based equating with other types of test formats. Third, careful attention should be paid to the generalization of these findings because of the specific conditions in these two simulation studies (as shown in **Tables 8**, **9**). For example, the discrimination parameter used in our studies are higher compared with other test equating studies. Future research should continue to investigate the performance of TRT-based models in other equating contexts, such as the equating for the multidimensional tests (Kim et al., 2019).

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

FH designed the study, conducted the simulation study, and drafted the manuscript. ZL participated in designing the study and conducted the simulation study. JS conducted the literature review. YL and LY conducted the data analysis. MZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was supported by Guangzhou education scientific planning subject (Grant No. 1201411413).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.743362/full#supplementary-material

REFERENCES

Andersson, B. (2018). Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Appl. Psychol. Meas.* 42, 192–205. doi: 10.1177/0146621617721249

Babcock, B., and Hodge, K. J. (2019). Rasch versus classical equating in the context of small sample sizes. *Educ. Psychol. Meas.* 80, 1–23. doi: 10.1177/0013164419878483

Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika* 64, 153–168. doi: 10.1007/BF02294533

Cai, L. (2017). *flexMIRT: Flexible Multilevel Multidimensional Item Analysis and Test Scoring (Version 3.51) [Computer Software]*. Chapel Hill, NC: Vector Psychometric Group.

Cao, Y., Lu, R., and Tao, W. (2014). *Effect of Item Response Theory (IRT) Model Selection on Testlet-Based Test Equating* (Research Report No. RR-14-19). Princeton, NJ: Educational Testing Service. doi: 10.1002/ets2.12017

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *J. Educ. Meas.* 43, 145–168. doi: 10.1111/j.1745-3984.2006.00010.x

González, J., and Wiberg, M. (2017). *Applying Test Equating Methods Using R*. New York, NY: Springer.

Hanson, B. A., and Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common item equating design. *Appl. Psychol. Meas.* 26, 3–24. doi: 10.1177/0146621602026001001

Keller, L., Swaminathan, H., and Sireci, S. G. (2003). Evaluating scoring procedures for context-dependent item sets. *Appl. Meas. Educ.* 16, 207–222. doi: 10.1207/S15324818AME1603_3

Kim, S. Y., Lee, W. C., and Kolen, M. J. (2019). Simple-structure multidimensional item response theory equating for multidimensional tests. *Educ. Psychol. Meas.* 80, 1–35. doi: 10.1177/0013164419854208

Kolen, M., and Brennan, R. (2014). *Test Equating, Scaling, and Linking: Methods and Practices*, 3rd Edn. New York, NY: Springer.

Koziol, N. A. (2016). Parameter recovery and classification accuracy under conditions of testlet dependency: a comparison of the traditional 2PL, testlet, and bi-factor models. *Appl. Meas. Educ.* 29, 184–195. doi: 10.1080/08957347.2016.1171767

Lawrence, I. M. (1995). *Estimating Reliability for Tests Composed of Item Sets* (Research Report No. RR-95-18). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1995.tb01653.x

Lee, G., Kolen, M. J., Frisbie, D. A., and Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Appl. Psychol. Meas.* 25, 357–372. doi: 10.1177/01466210122032226

Lee, P., Joo, S. H., and Stark, S. (2016). Linking methods for the zinnes-griggs pairwise preference IRT model. *Appl. Psychol. Meas.* 41, 130–144. doi: 10.1177/0146621616675836

Li, Y. M., Bolt, D. M., and Fu, J. B. (2006). A comparison of alternative models for testlets. *Appl. Psychol. Meas.* 30, 3–21. doi: 10.1177/0146621605275414

Nering, M. L., and Ostini, R. (eds) (2010). *Handbook of Polytomous Item Response Theory Models*. Abingdon-on-Thames: Routledge.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rijmen, F. (2009). *Three Multidimensional Models for Testlet-Based Tests: Formal Relations and An Empirical Comparison* (Research Report No. RR-09-37). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2009.tb02194.x

Sireci, S. G., Tissen, D., and Wainer, H. (1991). On the reliability of testlet-based tests. *J. Educ. Meas.* 28, 237–247. doi: 10.1002/j.2333-8504.1991.tb01389.x

Tao, W., and Cao, Y. (2016). An extension of IRT-based equating to the dichotomous testlet response theory model. *Appl. Meas. Educ.* 29, 108–121. doi: 10.1080/08957347.2016.1138956

van der Linden, W. J. (ed.) (2016). *Handbook of Item Response Theory: Models*, Vol. 1. Boca Raton, FL: Chapman & Hall/CRC.

von Davier, A. A., and Wilson, C. (2007). IRT true-score test equating: a guide through assumptions and applications. *Educ. Psychol. Meas.* 67, 940–957. doi: 10.1177/0013164407301543

von Davier, M., and von Davier, A. (2011). "A general model for IRT scale linking and scale transformations," in *Statistical Models for Test Equating,*

Scaling, and Linking, Vol. 1, ed. A. von Davier (New York, NY: Springer), 225–242.

Wainer, H., and Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case for testlets. *J. Educ. Meas.* 24, 185–201. doi: 10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., and Wang, X. H. (2001). *Using A New Statistical Model for Testlets to Score TOEFL*. (Research Report No. RR-01-09). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2001.tb01851.x

Wainer, H., Bradlow, E. T., and Du, Z. (2000). "Testlet response theory: an analog for the 3PL model useful in testlet-based adaptive testing," in *Computerized Adaptive Testing: Theory and Practice*, eds W. J. van der Linden and G. A. Glas (Dordrecht: Springer), 245–269.

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. New York, NY: Cambridge University Press.

Wainer, H., and Wang, X. H. (2000). Using a new statistical model for testlets to score TOEFL. *J. Educ. Meas.* 37, 203–220. doi: 10.1111/j.1745-3984.2000.tb01083.x

Wang, W. C., and Wilson, M. (2005). The Rasch testlet model. *Appl. Psychol. Meas.* 29, 126–149. doi: 10.1177/0146621604271053

Wang, X., Bradlow, E. T., and Wainer, H. (2002). A general Bayesian model for testlets: theory and applications. *Appl. Psychol. Meas.* 26, 109–128. doi: 10.1002/j.2333-8504.2002.tb01869.x

Wang, X., Bradlow, E. T., and Wainer, H. (2004). *User's Guide for SCORIGHT (Version 3.0): A Computer Program for Scoring Tests Built of Testlets Including A Module for Covariate Analysis*. (Research Report No. RR-04-49). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2004.tb01976.x

Wingersky, M. S., Cook, L. L., and Eignor, D. R. (1987). *Specifying the Characteristics of Linking Items Used for Item Response Theory Item Calibration* (Research Report No. RR-87-24). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2330-8516.1987.tb00228.x

Zenisky, A. L., Hambleton, R. K., and Sireci, S. G. (2002). Identifcation and evaluation of local item dependencies in the medical college admissions test. *J. Educ. Meas.* 39, 291–309. doi: 10.1111/j.1745-3984.2002.tb01144.x

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Lang. Test.* 27, 119–140. doi: 10.1177/0265532209347363

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 | Frontiers in Psychology

ORIGINAL RESEARCH
published: 01 April 2022
doi: 10.3389/fpsyg.2022.807049

Development and Psychometric Evaluation of Family Caregivers' Hardiness Scale: A Sequential-Exploratory Mixed-Method Study

Lida Hosseini[1], Hamid Sharif Nia[2] and Mansoureh Ashghali Farahani[1]**

[1] School of Nursing & Midwifery, Iran University of Medical Sciences, Tehran, Iran, [2] School of Nursing and Midwifery, Mazandaran University of Medical Sciences, Sari, Iran

Objective: Caring for patients with Alzheimer's disease (AD) is a stressful situation and an overwhelming task for family caregivers. Therefore, these caregivers need to have their hardiness empowered to provide proper and appropriate care to these older adults. From the introduction of the concept of hardiness, few studies have been conducted to assess the hardiness of caregivers of patients with AD. Presumably, one reason for this knowledge gap is the lack of a proper scale to evaluate hardiness in this group. This study was conducted to develop a reliable and valid Family Caregivers' Hardiness Scale (FCHS) to measure this concept accurately among Iranian family caregivers sample.

Methods: This study is a cross-sectional study with a sequential-exploratory mixed-method approach. The concept of family caregivers' hardiness was clarified using deductive content analysis, and item pools were generated. In the psychometric step, the samples were 435 family caregivers with a mean age of 50.26 (SD ± 13.24), and the data were gathered *via* an online form questionnaire. In this step, the items of the FCHS were evaluated using face and content validity. Then, the factor structure was determined and confirmed using exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) followed by convergent and divergent validity, respectively. Finally, scale reliability, including stability, and internal consistency were evaluated.

Results: The finding revealed that FCHS consists of five factors, namely, "Religious Coping" (5 items), "Self-Management" (6 items), "Empathic Communication" (3 items), "Family Affective Commitment" (3 items), and "Purposeful Interaction" (4 items) that explained 58.72% of the total variance. The results of CFA showed a good model fit. Reliability showed acceptable internal consistency and stability.

Conclusion: Based on the results of the psychometric evaluation of the FCHS, turned out that the concept of hardiness in Iranian family caregivers is a multidimensional concept that is most focused on individual-cultural values, emotional family relationships, and social relationships. The designed scale also has acceptable validity and reliability features that can be used in future studies to measure this concept in family caregivers.

Keywords: family caregivers, Alzheimer, hardiness, validity, psychometric, scale

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Katerina M. Marcoulides,
University of Minnesota Twin Cities, United States
Álvaro Postigo,
University of Oviedo, Spain

****Correspondence:***
Hamid Sharif Nia
pegadis@yahoo.com
Mansoureh Ashghali Farahani
m_negar110@yahoo.com

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

Received: *19 November 2021*
Accepted: *07 February 2022*
Published: *01 April 2022*

Citation:
Hosseini L, Sharif Nia H and Ashghali Farahani M (2022) Development and Psychometric Evaluation of Family Caregivers' Hardiness Scale: A Sequential-Exploratory Mixed-Method Study. Front. Psychol. 13:807049. doi: 10.3389/fpsyg.2022.807049

INTRODUCTION

Aging has become one of the greatest concerns around the world due to increasing life expectancy and decreasing mortality (Santos da Silva et al., 2018). Based on the WHO reports, in 2019, 703 million people aged 65 years and older worldwide, and it will reach 1.5 billion people by 2030. This increase in developing countries such as Iran will occur faster than in developed countries (World Health Organization [WHO], 2020).

Aging is a natural and inevitable process of life and is associated with a series of physical, cognitive, and emotional changes (Pashaki et al., 2015). Alzheimer's disease (AD) is one of the most common types of cognitive disorder that affects the memory, thinking, and behavior of older adults and reduces the person's ability to live independently (Santos da Silva et al., 2018). The Alzheimer's Disease International (ADI) Federation estimates 35.6 million people live with AD worldwide, and it will double every 5 years after the age of 65 years (Alzheimer's Disease International [ADI], 2017; Trevisan et al., 2019). Since older adults with AD are limited in performing their activities of daily living, they need to be supported by a formal or informal caregiver (Santos da Silva et al., 2018). Due to the interdependence between family members, declining household incomes especially in developing countries such as Iran, the lack of formal support systems, more than 81% of these patients are in need of care by family caregivers (Sharifi et al., 2016). Family caregivers are considered informal caregivers and lack training; these individuals do not receive any reimbursement for their services (Lynch et al., 2018).

The caregiver burden for family caregivers of patients with AD is heavy work, and caring for patients with AD is stressful and can become overwhelming for family caregivers. As the severity of the disease increases, it affects all aspects of these caregivers' lives and can produce many acute and chronic physical and emotional problems for family caregivers (Armstrong et al., 2019). Thus, caregivers can be considered "invisible secondary patients" (Ashrafizadeh et al., 2021). Previous studies have shown that depression, anxiety, stress, and burnout are the most common sequela of caring for family caregivers of patients with AD. So that, more than 80% of caregivers suffer from stress and burnout, 30–40% suffer from depression, and 44% suffer from anxiety (Baharudin et al., 2019; Fujihara et al., 2019). Therefore, for these caregivers to be able to adapt properly to the situation and not suffer from the negative side effects, they need the ability, competence, and skills to adapt to the situation (Lynch et al., 2018). According to Hooker et al., personal characteristics such as hardiness can be a major factor in changing the care experience when caring for patients with AD and increase the caregivers' ability for positive coping with these stresses (Hooker et al., 1998).

Background

Hardiness was first proposed by Kobasa (1979). It is one of these effective personal characteristics which makes sense in the face of stressful situations and is considered as a moderating variable in the relationship between stress and its physical and psychological effects (Abdollahi et al., 2018). According to Kobasa, hardiness is a combination of attitudes and beliefs that motivate a person to do hard and strategic work in the face of stressful and difficult situations and can turn adversity into an opportunity for growth (Maddi, 2002). Accordingly, this concept consists of three components, namely, commitment, control, and challenge (Kobasa, 1979). Commitment refers to a tendency to engage in life's activities and to have a real interest and curiosity about the world around them. Control refers to the belief that individuals can influence the events of their lives; and finally, challenge points to the belief that change, rather than stability, is a natural part of life, which creates opportunities for personal growth rather than threatening security (Maddi, 2002). Studies show the positive effect of hardiness on health and performance in different groups such as college students, cadets, nursing students, and managers in different stressful situations (Kelly et al., 2014; Abdollahi et al., 2018; Tho, 2019). One meta-analytic review showed that hardy individuals are likely to have more life satisfaction, a better job or school performance, more optimism, greater self-esteem, and a sense of coherence as well as higher mental health; but individuals with low hardiness experience more negative effects from stressful situations such as depression and anxiety (Eschleman et al., 2010).

Since caring for patients with AD is a unique and stressful situation for family caregivers, to provide proper and appropriate care to these patients, these caregivers need to have the hardiness trait in order to be empowered. As a moderating factor, hardiness can prevent problems for the caregivers such as fatigue, burnout, depression, sleep disorders, and reduced quality of life. Hardiness can also prevent the patients from neglect, abuse, poor quality care, ignoring vital needs, and aggravation of the disease (Clark, 2002; DiBartolo and Soeken, 2003). It is noteworthy that since the introduction of the concept of hardiness, few studies have assessed the hardiness of caregivers of patients with AD. Presumably, one reason for this knowledge gap is the lack of a proper scale to evaluate hardiness in this group. Several questionnaires have been developed for measuring hardiness in different groups such as students (Benishek and Lopez, 2001), bereaved parents (Lang et al., 2003), and employees (Moreno-Jiménez et al., 2014). However, the caregiving for patients with AD is completely different from the previous studies about the role of hardiness.

Therefore, considering that the Iran population is aging, AD is an age-related phenomenon, and that patients with AD are mostly cared for by family caregivers, therefore, Iran will need to prepare hardy family caregivers. Furthermore, since hardiness can be taught to individuals, nurses and therapists will be able to design appropriate interventions to improve their hardiness and thus improve care and reduce complications. Knowing the level of the caregiver's hardiness or evaluating the effectiveness of interventions requires an accurate scale. Thus, this study was conducted to clarify the concept of hardiness in family caregivers of patients with AD and then develop a reliable and valid scale to measure this concept accurately.

MATERIALS AND METHODS

Design

This is a cross-sectional study to evaluate the psychometrics of the Family Caregivers' Hardiness Scale (FCHS) from July 2020 to October 2021 in family caregivers of patients with AD. It was performed in two stages: (1) qualitative by directed content analysis approach to generate items and (2) quantitative approach to assess the psychometric properties of the developed scale.

Qualitative Study and Item Generation

The purpose of this stage was to clarify the family caregivers' hardiness concept and make an item pool for designing the target scale. For this purpose, based on the Kobasa's model of hardiness, the deductive directed content analysis by Elo and Kyngäs (2008) was used to clarify the concept of the family caregivers' hardiness in caring for patients with AD. The related structures were identified, and the items were produced in two steps: reviewing the literature and examining the experiences and perceptions of the participants through interviews. The deductive-directed content analysis includes three phases, namely, preparation, organization, and reporting.

First Step: A Review of the Literature

Electronic databases such as PubMed, Scopus, ISI Web of Science, and Persian databases such as Magiran, SID, and Iran Medex were searched using the keywords "hardiness," "personality hardiness," "hardy personality," "caregiver hardiness," "caregivers," "family caregivers," "non-professional caregivers," "spouse caregivers," "dementia," and "Alzheimer" with no time limit. Studies with the following inclusion criteria were selected: relevance of the study, access to the full text of the article, and English and Persian language. In this search, duplicate and irrelevant articles, studies published in non-Persian and non-English languages, and short articles such as the editorial and commentarial materials were excluded. In the initial search, a total of 3,560 English articles and 430 Persian articles were obtained. After applying the inclusion and exclusion criteria, 23 articles were entered the analysis stage to extract initial codes. In the preparation phase, the text of each article was read several times by the researcher (L.H) as a unit of analysis to immerse in the data and to provide key points and clear descriptions of each aspect of the hardiness concepts based on the Kobasa hardiness model. Then, in the organizing phase, the researchers formed an unconstrained matrix derived from the Kobasa Hardiness Model. Initial codes (n = 198) were classified as categories derived from the dimension of hardiness (i.e., main categories of commitment, control, and challenge and two new main categories: connection and culture). The choice of these names for the main categories was based on the hardiness concept.

Second Step: An Interview With Participants

Participants

To deeply understand the family caregivers' hardiness concept, 14 family caregivers with a mean age of 54.57 years were selected through purposeful sampling with maximum variety and also snowball sampling from November 2020 to February 2021. Personal characteristics were as follows: nine daughters, two sons, and three spouses. Ten participants were married, three were unmarried, and one of them was a widow. Eight participants had an academic education, and six had diploma.

Procedure

In-depth and semi-structured interviews (30–90 min) were conducted with each participant using a combination of model-derived questions and open-ended questions. Immediately after the end of each interview, the recorded material was transcribed word by word. In the preparation phase, the researcher (L.H) listened to the recorded statements and read the written interview several times to gain an in-depth understanding of the participants' feelings and experiences and then analyzed it using MAXQDA software version 10. In the organizing phase, similar to the review of the literature step, the researchers formed an unconstrained matrix derived from the Kobasa hardiness model, and a total of 1,604 initial codes were extracted, leaving 606 initial codes after deleting duplicates and overlapping cases. These were classified as categories derived from the dimension of hardiness (i.e., main categories of commitment, control, and challenge and two new main categories: connection and culture). Finally, in reporting phase, the results of both steps were combined (Hosseini et al., 2021). Also, all stages of directed content analysis and the findings obtained in this study were reported. The quality of findings was assessed by Lincoln and Goba's criteria such as credibility, dependability, confirmability, and transferability (Lincoln and Guba, 1986). Finally, based on the result of the concept analysis and the extracted codes, an item pool (656) was developed. Later, during frequent meetings of the research team, writing and grammar and also overlapping and similarity of items were checked, and some items were merged or deleted. Thus, the total number of items was reduced from 656 to 97 and then to 54 items. Therefore, at this stage, the 54-item FCHS was developed to be evaluated for psychometric properties.

Quantitative Study and Item Reduction

During this stage, face, content, and construct validity, as well as reliability, were used to evaluate the psychometric properties of the FCHS using a five-point Likert response scale, i.e., 5 (always) to 1 (never). The sample size of each stage was different, and it was explained separately in each stage.

Face Validity

Face validity was evaluated with qualitative and quantitative approaches. In the qualitative approach, the scale was sent to 11 family caregivers who were asked to assess the scale in terms of difficulty, relevancy, and ambiguity. All items were understandable to the participants. In the quantitative approach, we asked the same 11 family caregivers to assess the items in terms of suitability using a five-point Likert scale (5 = it is completely suitable, 4 = it is suitable, 3 = it is almost suitable, 2 = it is less suitable, and 1 = it is not suitable at all). The impact score was calculated with the formula as follows: impact score = frequency (%) $\times$ suitability. A score of $>$1.5 was considered acceptable (Ebadi et al., 2020).

Content Validity

The content validity of the FCHS was evaluated by the qualitative and quantitative approaches. In the qualitative approach, the scale was sent to 12 experts in nursing, psychology and the development of the instrument to evaluate the items in terms of grammar and wording, item allocation, and scaling. During this process, some items were modified by their feedback. In the quantitative approach, the content validity of the scale was evaluated by content validity ratio (CVR) and modified kappa coefficient (K) to ensure that the scale measures the construct of interest. In CVR, 12 experts evaluated the essentiality of FCHS in a three-point Likert scale (1 = not essential, 2 = useful but not essential, and 3 = essential). The CVR was accounted by the formula as follows: [ne – (N/2)]/(N/2), where "ne" is the number of experts who rate the items as "Essential" and *N* is the total number of experts. The result was interpreted using the Lawshe rule. The minimum acceptable CVR score was 0.56 (Lawshe, 1975). To assess K to the elimination of chance effect for each item, 11 experts evaluated the 38-item scale in terms of relevancy by the dichotomous response: (4 = relevant, 1 = irrelevant). An excellent value of kappa was considered as >0.75 (Ebadi et al., 2020).

Item Analysis

Before examining the construct validity, an item analysis was conducted to identify possible problems of items by computing the corrected item-total correlation. In this step, 32 family caregivers with a mean age of 52.02 ± 13.91 years were selected using convenience sampling. They completed the online form of FCHS. We considered the correlation coefficient between items lower than 0.32 or above 0.9 as criteria for removing items (Ebadi et al., 2020).

Construct Validity

Participations and Samples

The sample consisted of Iranian family caregivers of patients with AD. The inclusion criteria to participate in this study were as follows: be the family member, relatives, and friends of the patient (informal caregivers) and providing care for the patient, agreed to participate in this study, and able to use social networks such as Telegram and WhatsApp. Based on the Rule of Thumb that considers 200 participants as the adequate sample size (MacCallum et al., 1999), 435 family caregivers were recruited into this phase for two steps: 210 for evaluating exploratory factor analysis (EFA) and 225 for evaluating confirmatory factor analysis (CFA). The participants were selected using convenience sampling through social groups related to the family caregivers of patients with AD and through the introduction of people. During this phase, data were gathered online. For this purpose, the online questionnaire was created *via* Google Form, and its URL link was sent by email or social networking applications such as Telegram channel or WhatsApp for participants.

Measures

The questionnaire used in this step included two sections. The first section was related to the demographic characteristics such as age, gender, marital status, education level, employment status, lifestyle, relationship with the patient, average hours of care per day (h), and duration of the disease (year). The second section was FCHS with 32 items to the measuring of the family caregiver's hardiness concept with a five-point Likert scale response (1 = never to 5 = always). The details of the production phases of FCHS are shown in **Figure 1**.

The construct validity of this scale was evaluated by EFA and CFA. The EFA was assessed through the maximum-likelihood method with Promax Rotation using SPSS/AMOS$_{26}$. Furthermore, the Kaiser-Meyer-Olkin (KMO) and Bartlett's tests were used to estimate sample adequacy and suitability. KMO values higher than 0.9 were interpreted as excellent (Pahlevan and Sharif, 2021). Horn's parallel analysis and exploratory graph analysis were used for extracting factor structure using SPSS R-Menu$_{2.0}$. Horn's parallel analysis method is found to have consistent results to determine the accurate number of factors and the original scale. Horn's parallel analysis creates eigenvalues that take into account the sampling error inherent in the dataset by creating a random score matrix of exactly the same rank and type of the variables we have in our dataset. The actual matrix values are then compared with the randomly generated matrix. The numbers of components, after successive iterations, that account for more variance than the components derived from the random data are taken as the correct number of factors to extract (Pahlevan and Sharif, 2021). Factor loading of almost 0.3 was considered to determine the presence of an item in a latent factor, and items with communalities < 0.2 were excluded from EFA. Factor loading was estimated using the following formula: CV = 5.152 ÷ $\sqrt{}$ (n – 2), where CV is the number of extractable factors, and N is the sample size (Pahlevan and Sharif, 2021). Then, the factor structure determined by EFA was assessed by CFA. The CFA was performed using the maximum-likelihood method and the most common goodness-of-fit indices such as chi-square (χ^2) test, chi-square/degree-of-freedom ratio (χ^2/df) < 3, Comparative Fit Index (CFI) > 0.90, Incremental Fit Index (IFI) > 0.90, Tucker-Lewis Index (TLI) > 0.90, Parsimonious Normed Fit Index (PNFI) > 0.50, Parsimonious Comparative Fit Index (PCFI) > 0.50, and root mean square error of approximation (RMSEA) < 0.08 using SPSS/AMOS$_{26}$ (Hu and Bentler, 1999; Pahlevan and Sharif, 2021).

Convergent and Discriminant Validity

The convergent and discriminant validity of the extracted factors was evaluated using Fornell-Larcker criteria using JASP$_{15.0.0}$ as follows: (a) average variance extracted (AVE), (b) maximum shared squared variance (MSV), and (c) composite reliability (CR). The AVE > 0.5 and (b) CR greater than AVE was considered as the minimum requirements of convergent validity. Also, MSV less than AVE for each construct was considered the minimum requirement of the discriminant validity (Pahlevan and Sharif, 2021). In this study, the discriminant validity was assessed by a new approach developed by Heseler as Heterotrait-Monotrait Ratio (HTMT) matrix in which, to achieve discriminant validity,

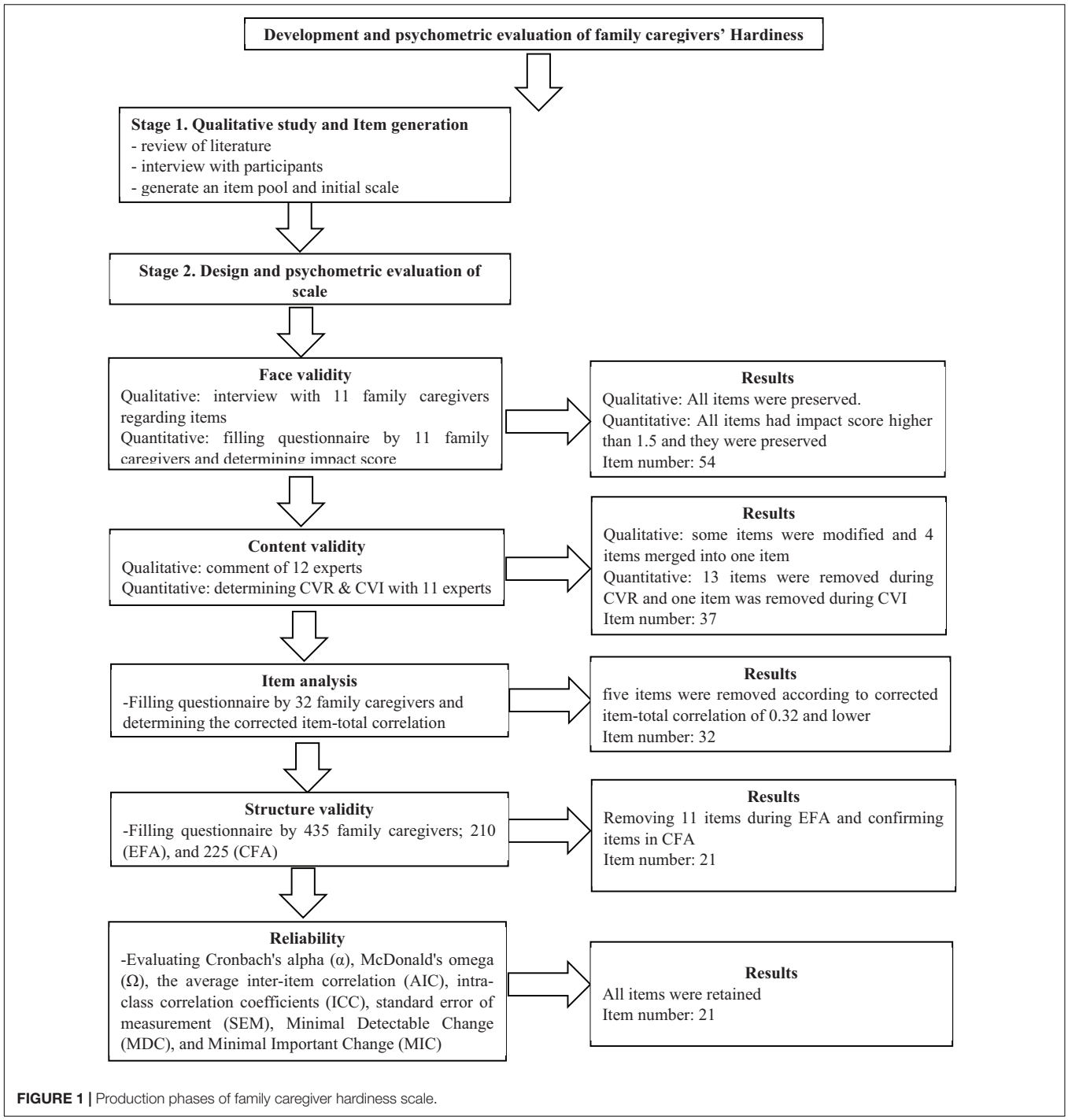

FIGURE 1 | Production phases of family caregiver hardiness scale.

all values in the HTMT matrix should be less than 0.85 (Henseler et al., 2015).

Reliability

Reliability was evaluated using internal consistency, stability, and absolute reliability approaches using $JASP_{15.0.0}$. The internal consistency was evaluated using Cronbach's alpha (α), McDonald's omega (Ω), and the average inter-item correlation (AIC). Coefficient's α and Ω values were > 0.7, and the AIC of 0.2–0.4 was considered as an acceptable internal consistency (Sharif Nia et al., 2021). Also, CR and maximum reliability (Max H reliability) > 0.7 were used to evaluate the reliability of the construct in the structural education model (Sharif Nia et al., 2019).

The stability was evaluated by counting the intraclass correlation coefficients (ICC) of the FCHS with a two-way random effects model. For this purpose, we used the test-retest method with a 2-week interval in 15 family caregivers. The

ICC value > 0.8 is considered an acceptable value of stability (Pahlevan and Sharif, 2021).

Furthermore, the absolute reliability was evaluated using standard error of measurement (SEM) by the following formula: ($SEM = SD_{Pooled} \times \sqrt{1 - ICC}$) (Pahlevan and Sharif, 2021).

Finally, the responsiveness was assessed using the minimal detectable change (MDC) by using the following formula: $MDC95\% = SEM \times \sqrt{2} \times 1.96$ and the minimal important change (MIC) by using the following formula: MIC = 0.5 × SD of the Δ score. To interpret the MIC, it is necessary to calculate the limit of agreement (LOA). The LOA was calculated based on the following formula: LOA = $d \pm 1.96 \times$ SD difference. If the MIC is smaller than the MDC or the MIC is not within LOA, the scale is responsive. Also, interpretability was assessed by evaluating ceiling and floor effect and MDC (Ebadi et al., 2020).

Multivariate Normality and Outliers

The normal distribution of data was evaluated in two ways, namely, univariate and multivariate distributions. Univariate normal distribution was evaluated using skewness (±3) and kurtosis (±7), and multivariate normality distribution was assessed by Mardia's coefficient > 8. The data were evaluated for the outlier in two ways, namely, univariate and multivariate outliers. The univariate outlier was assessed through distribution charts, and the multivariate outlier was assessed through Mahalanobis distance $p < 0.001$ (Pahlevan and Sharif, 2021).

Ethical Consideration

The Iran University of Medical Sciences Research Ethics Committee approved this study (IR. IUMS. REC.1398.1229). In the beginning of each interview, the purpose of the interview was explained to the participants, and they were asked to provide written permission and informed consent to audio record their answers to questions. In addition, they were reassured that participation in the study was voluntary. Participants were assured that their information was confidential.

RESULTS

Item Generation

The results of the review of literature and interview with participants were combined. Based on the results of this phase, the concept of family caregivers' hardiness of patients with AD had five dimensions, namely, commitment, control, challenge, connection, and culture. The item pool with 656 items was generated using initial codes. Out of which 54 items were selected as items of the FCHS.

Item Reduction

In the face validity step, the score of all items was above 1.5, and they were found to be suitable. During the assessment of content validity, in the qualitative approach, four items merged into one item according to expert panel suggestion. In quantitative approaches, the CVR of 13 items were < 0.56, and they were removed, and according to the results of kappa value, the kappa value of one item was < 0.75, and it was removed (4 items from the first dimension, 10 items from the second dimension, 1 item from the fourth dimension, and 2 items from the fifth dimension). Therefore, 17 items were removed, and the total number of the FCHS was reduced from 54 to 37 items. During the item analysis step, five items (i.e., items 12, 16, 19, 27, and 33) were also removed, because they were corrected, the item-total correlation of 0.32 and lower and the final FCHS with 32 items were entered into the factor analysis step.

Sociodemographic Profile of Participants

In total, 435 family caregivers with a mean age of 50.26 years (SD = 13.24) participated in this study. The number of women (50.6%) and men (49.4%) were almost equal. Most of them were married (68.7%) and daughters of patients (52.9%). The details of the sociodemographic profile of participants were shown in **Table 1**.

In the construct validity step, based on the results of KMO (0.935) and Bartlett's value 2132.372 ($p < 0.001$), the sample was adequate and suitable. In this step, 11 items (items 4, 6, 8, 10, 13, 20, 23, 24, 25, 27, and 30) that were removed as the communality values of them were less than 0.2, and the factor loadings were less than 0.3, and after Promax Rotation, five-factors (21 items totally) such as "Religious Coping" (5 items), "Self-Management" (6 items), "Empathic Communication" (3 items), "Family Affective Commitment" (3 items), and "Purposeful Interaction" (4 items) were extracted. These factors explained, respectively, 16.37, 15.83, 8.96, 8.51, 9.11, and 58.72% of the total variance of family

TABLE 1 | Demographic characteristics of participants (n = 435).

Variables		*N* (%)
Age		50.26 ± 13.24
Gender	Female	220 (50.6)
	Male	215 (49.4)
Marital status	Single	92 (21.1)
	Married	299 (68.7)
	Divorced	14 (3.2)
	Widow	30 (6.9)
Education level	Illiterate	11 (2.5)
	Less than diploma	30 (6.9)
	Diploma	200 (46)
	Academic	194 (44.6)
Employment	Unemployed	42 (9.7)
	Employed	161 (37)
	Housewife	146 (33.6)
	Retiered	24 (5.5)
	Free	62 (14.3)
Lifestyle	Independent	262 (60.2)
	With patients	173 (39.8)
Relationship with the patient	Daughter	230 (52.9)
	Son	57 (13.1)
	Wife/midwife	57 (13.1)
	Friend	34 (7.8)
	Relative	57 (13.1)
Average hours of care per day (hour)		7.51 ± 5.51
Duration of the disease (year)		4.65 ± 2.52

TABLE 2 | The result of EFA on the five factors of FCHS (N = 210).

Factors	Q_n. Item	Factor loading	h^{2*}	M (SD)	Skew (kurtosis)	λ	%Variance
Religious coping	**31.** Believing in God's help in trouble will make me stronger in the face of adversity.	0.938	0.811	4.06 (1.16)	−1.2 (0.70)	3.43	16.37
	32. Prayer and communion with God make me hardy against the pressure of care.	0.898	0.653	3.96 (1.33)	−1.0 (−0.27)		
	29. The spiritual value of patient care makes it easier for me to endure care problems.	0.890	0.776	3.68 (1.37)	−0.8 (−0.27)		
	22. Caring for the patient as a spiritual opportunity strengthens me.	0.776	0.674	3.59 (1.34)	−0.9 (0.03)		
	21. I see caring for the patient as an opportunity to repay efforts and pay my homage to the patient.	0.598	0.420	4.09 (1.14)	−1.5 (1.06)		
Self-management	**15.** By changing my mind in difficult times, I try to bear the pressure of care.	0.878	0.673	3.53 (1.07)	−0.4 (−0.47)	3.32	15.83
	17. With care management, I endure problems.	0.856	0.714	3.90 (0.89)	−0.8 (0.35)		
	14. Recalling my own abilities, I try to bear the pressure of care.	0.798	0.571	3.71 (0.88)	−0.5 (−0.58)		
	16. I constantly remind myself that enduring the hardships of caring is part of my job.	0.695	0.492	3.75 (1.31)	−1.0 (0.61)		
	19. Positive thinking helps me not to give in to difficult situations.	0.598	0.709	3.56 (1.41)	−0.7 (0.14)		
	18. Patience in the face of problems for my patient makes the situation bearable.	0.588	0.582	3.75 (0.84)	−0.7 (−0.16)		
Empathic communication	**7.** Understanding the involuntary nature of the patient's problems makes it easier to endure hardships.	0.896	0.494	3.68 (1.06)	−1.0 (1.11)	1.88	8.96
	5. Accepting the patient's condition makes the difficulty of caring tolerable for me	0.778	0.646	4.28 (0.28)	−0.9 (0.70)		
	9. Creating a sense of satisfaction in patient, makes the care easier for me.	0.689	0.502	3.96 (1.17)	−1.0 (0.39)		
Family affective commitment	**1.** My interest in my family causes me; To endure the hardships of care.	0.850	0.605	4.56 (0.71)	−1.4 (1.60)	1.78	8.51
	3. Love for my patient makes me endure the hardships of caring.	0.762	0.616	4.40 (0.83)	−1.1 (0.61)		
	2. I am responsible to my family.	0.697	0.413	4.68 (0.64)	−1.8 (1.31)		
Purposeful interaction	**28.** Talking to a doctor or nurse about a patient's problems makes it easier for me to bear the pressure of care.	0.787	0.422	3.68 (1.02)	−0.4 (0.68)	1.91	9.11
	11. Gaining information about the disease through different methods (cyberspace, books, brochures, and treatment team) increases my ability to care.	0.692	0.457	3.68 (0.99)	−0.4 (0.77)		
	12. Sharing and exchanging ideas with family members makes it easier for me to endure problems.	0.685	0.472	3.87 (0.79)	−0.6 (−0.23)		
	26. Associating with friends and acquaintances makes the burden of care bearable for me.	0.589	0.244	3.81 (1.14)	−0.4 (−0.51)		

h^2, Communalities; λ, Eigenvalue.

caregivers' hardiness. The details of factor analysis results are shown in **Table 2** and **Figures 2**, **3**.

In the next step of construct validity, the model was tested by CFA. The results showed all of the model fit indices were in the acceptable range and showed the model of family caregivers' hardiness is fit (**Figure 4**). For example, the chi-square model fit index was 311.314 ($p < 0.001$), CMIN/DF was 1.759, RMSEA was 0.065. The results of the other model fit indices are shown in **Table 3**.

The first four factors of the scale which had convergent validity based on AVE, MSV, and CR results were used to assess convergent, discriminant validity. All items had discriminant validity. Furthermore, the results of HTMT showed that there are no warnings for discriminant validity (**Tables 4**, **5**).

The results of Cronbach's alpha, McDonald's omega, and AIC for five factors were greater than 0.7 and 0.4, respectively, and the internal consistency of the scale was acceptable. In addition, the scale had a strong coefficient based on the results of CR and Max H reliability (**Table 4**). Finally, the stability of scale was strong based on the overall ICC result (0.903, 95% CI: 0.719–0.967) (**Table 6**). Absolute reliability based on SEM results was 2.89. This value indicates that the scale score in a person varies ± 2.89 in repeated tests. Based on the results of MDC, MIC, LOA, and ceiling and floor effect, this scale had responsiveness. In addition, the results of the floor and ceiling effects showed that the items are free of these effects and the scale has interpretability (**Table 6**).

DISCUSSION

The results of this study indicated that family caregivers' hardiness concept has five dimensions, namely, commitment, control, challenge, connection, and culture of Iranian caregivers. Therefore, FCHS is a valid and reliable scale for assessing this

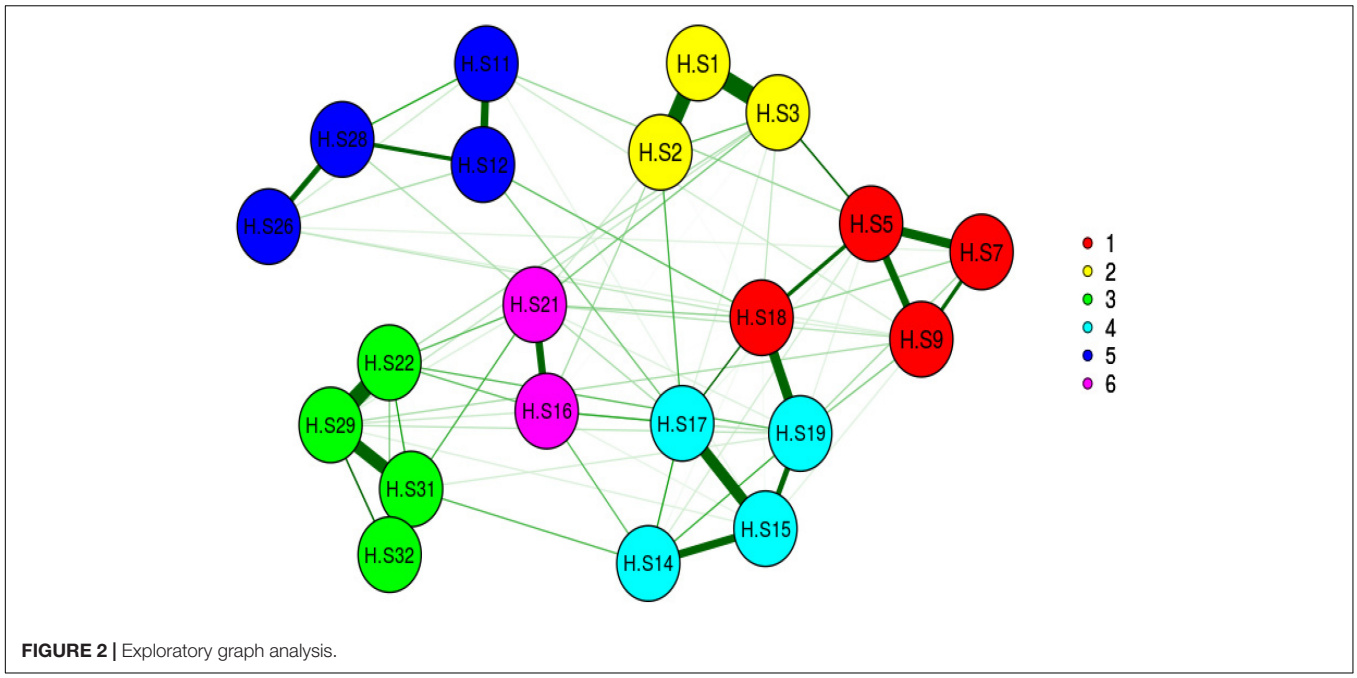

FIGURE 2 | Exploratory graph analysis.

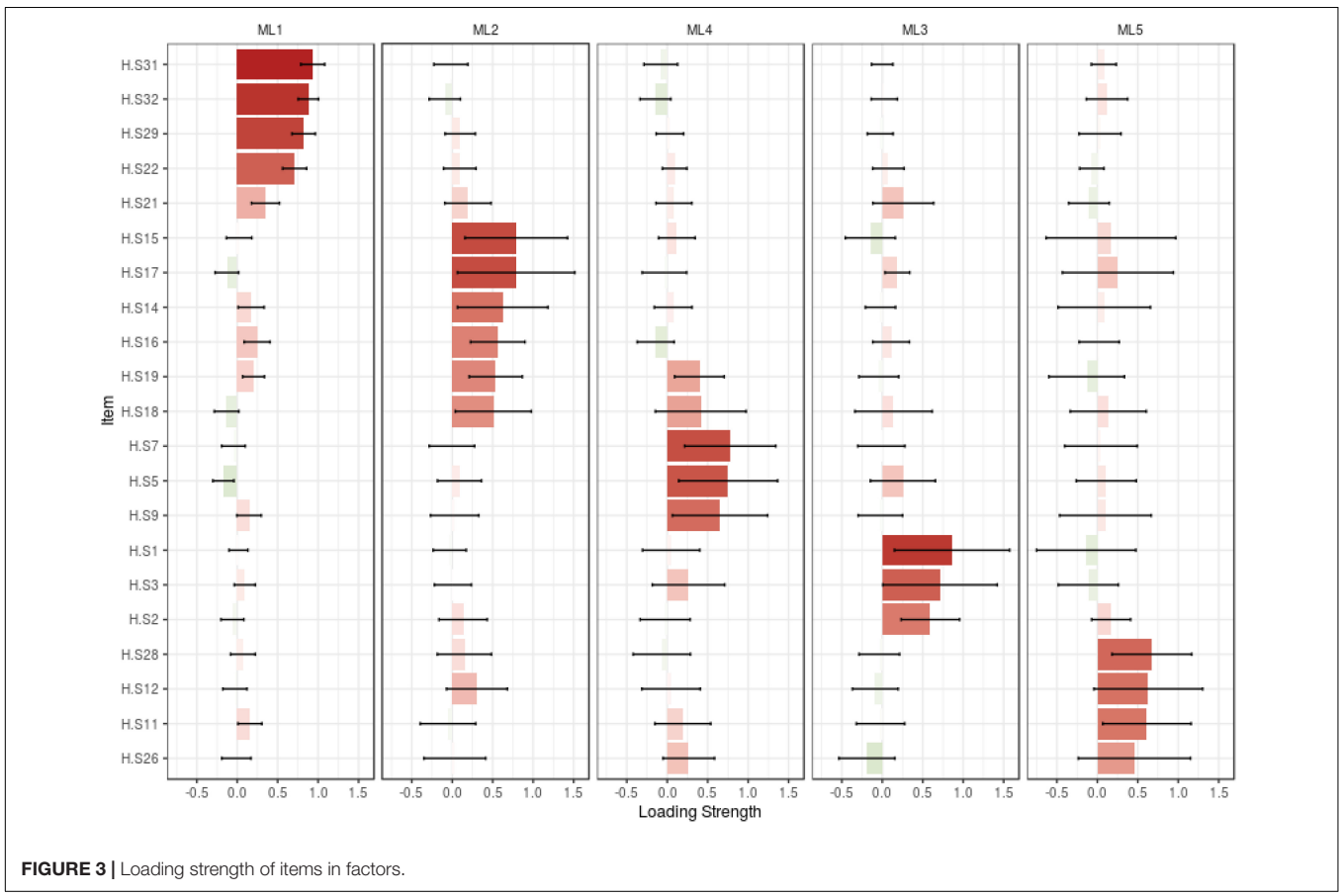

FIGURE 3 | Loading strength of items in factors.

concept in family caregivers of patients with AD. This scale includes 21 items and five factors, namely, religious coping, self-management, empathic communication, family affective commitment, and purposeful interaction that explained 58.72% of the total variance of this concept. The FCHS model obtained with EFA was confirmed with CFA. As the results of convergent

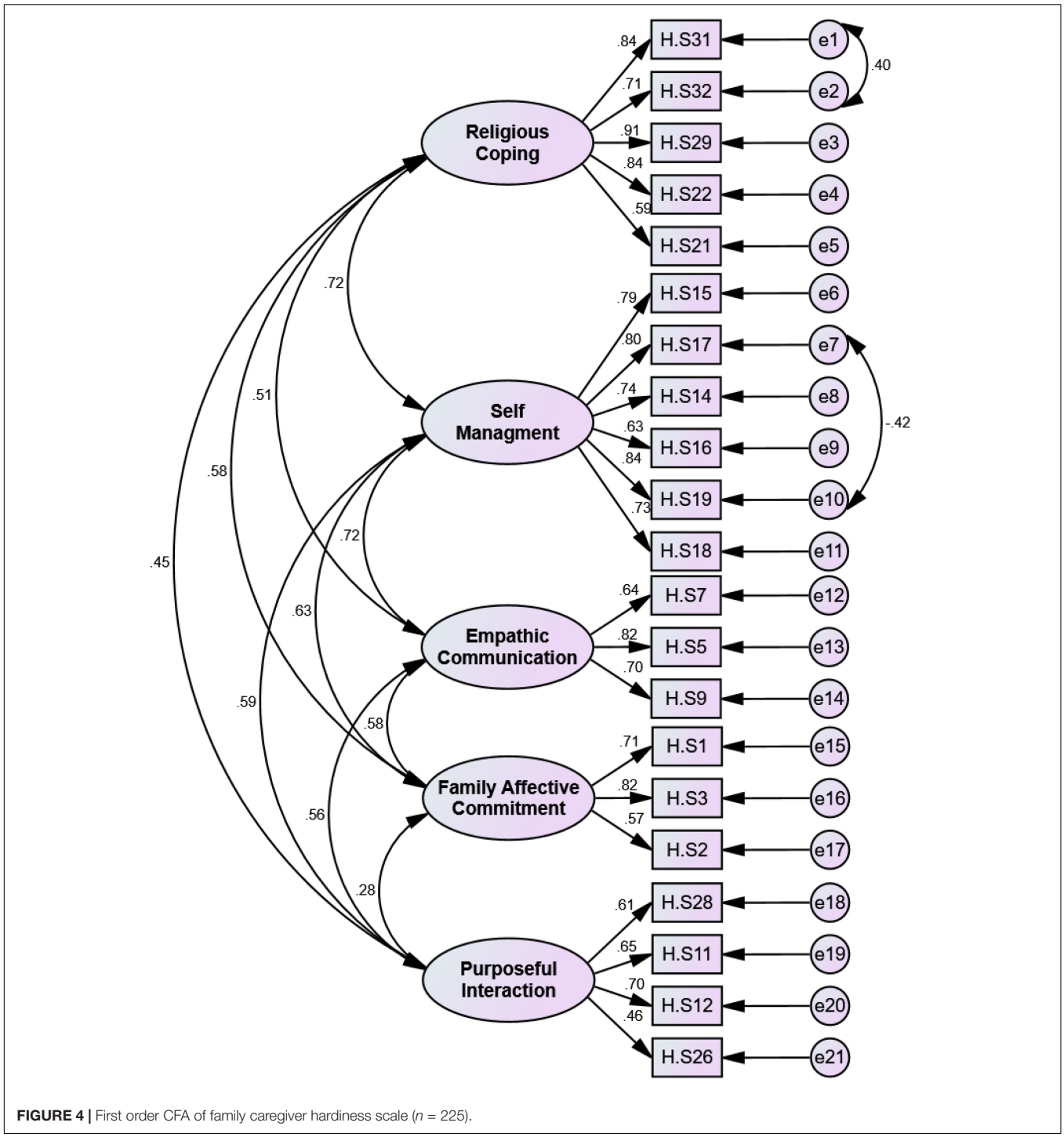

FIGURE 4 | First order CFA of family caregiver hardiness scale (n = 225).

and discriminant validity showed that the factors of this scale correlate with total scale, while they have a low correlation with each other. Therefore, the five factors of this scale are independent.

Since one of the main goals of the factor analysis is to maximize variance, in this study, the variance was 58.72% that factors one and two explained the greatest values of 16.37 and 15.83%, respectively. Among the scales designed to measure the concept of hardiness, regardless of the factor extraction method, two scales explained variance more than FCHS. The Children's Hardiness Scale (CHS) explained 65.75% (Soheili et al., 2021), and graduate students' academic hardiness (GSAH) explained 61.87% (Cheng et al., 2019).

Furthermore, this scale had excellent internal consistency based on the results of Cronbach's alpha, AIC, and McDonald's omega. It is noteworthy that one of the advantages of this scale

TABLE 3 | Factors adjustment indexes obtained in exploratory factor analysis of the FCHS (n = 225).

	1 Factor	2 Factor	3 Factor	4 Factor	5 Factor (Final model)
CMIN	6.797	90.868	141.204	203.053	311.314
df	3	41	72	111	177
P value	0.079	<0.001	<0.001	<0.001	<0.001
CMIN/DF	2.266	2.216	1.961	1.825	1.759
RMSEA	0.083	0.082	0.073	0.068	0.065
PNFI	0.706	0.713	0.720	0.726	0.719
PCFI	0.708	0.715	0.754	0.772	0.784
TLI	0.978	0.946	0.941	0.934	0.916
IFI	0.993	0.960	0.954	0.947	0.931
CFI	0.993	0.960	0.953	0.946	0.930

TABLE 4 | The indices of the convergent, discriminant validity, and internal consistency of FCHS OF CFA (N = 225).

	CR	AVE	MSV	MaxR (H)	Alpha [CI95%]	Omega	AIC
Religious coping	0.889	0.620	0.520	0.917	0.889	0.900	0.615
Self- management	0.890	0.575	0.520	0.898	0.880	0.882	0.557
Empathic communication	0.767	0.525	0.513	0.788	0.764	0.766	0.522
Family affective commitment	0.749	0.504	0.396	0.783	0.749	0.773	0.502
Purposeful interaction	0.699	0.372	0.351	0.716	0.691	0.692	0.364

is having strong stability based on the value of ICC. Another advantage of this study was the evaluation of measurement error, responsiveness, and interpretation of FCHS. So that the results showed, FCHS has the minimum amount of SEM, responsiveness, and interpretability. SEM indicates the accuracy of the measurement for each individual, and the smaller value of it is important. Responsiveness demonstrates the ability of a scale to show changes in a person's situation over a period. Finally, the interpretability shows the ability of the scale to show the meaningfulness of changes. These features are an important and required domain of the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) CHECKLIST (Terwee et al., 2007) that were not reported in the previous studies of the psychometric properties about hardiness.

The FCHS has five factors, namely, "religious coping," "self-management," "empathic communication," "family affective commitment," and "purposeful interaction." The first factor of FCHS was labeled "religious coping." It includes five items that explained 16.37% of the total variance. The religious coping concept is defined as using religious beliefs or behaviors to facilitate problem-solving to prevent or reduce the negative emotional consequences of stressful living conditions (Koenig et al., 1998). In this scale, religious coping was defined as the caregiver's ability to use religious and spiritual behaviors and beliefs to cope with the stresses of caring for a patient with AD. It is noteworthy that Mund in 2017 proposed culture as one of the five dimensions of hardiness concept (Mund, 2017); because based on the finding of previous studies, Mund had suggested that a strong background of culture had contributed to the formation of personality and coping strategies. The Iranian culture has been associated with religion and spirituality, and it helps people deal with stressful situations (Abdollahi and Abu Talib, 2015). Our study shows that religion and spirituality had the greatest impact on the hardiness of family caregivers of patients with AD. Therefore, the findings of our study reinforced Mund's suggestion as an introduction to the fifth component of hardiness.

The second extracted factor was "self-management" with 6 items. In line with the definitions provided for self-management (Barlow et al., 2002), this scale refers to self-management as the psychological mechanisms used to cope with the stresses of caring and to overcome difficult situations including positive thinking, self-remembering and self-emphasis, and patience with the individual to handle their emotions. This factor is related to the control component of hardiness (Kobasa, 1979). Furthermore, the meaning of the self-management factor is in line with the control of affect in the academic hardiness scale (Weigold et al., 2016) studied in GSAH (Cheng et al., 2019), because control of affect also assesses a person's ability to handle his/her emotions related to academic issues. Since caring for patients with AD has a more psychological burden for caregivers (Fujihara et al., 2019), having the ability to manage this burden is important, and based on the results of this study, self-management was recognized as the second most effective factor.

The third factor extracted was labeled "empathic communication" with 3 items. Empathetic communication is defined as "a two-step process involving: (1) an in-depth understanding of the other person's problem or feelings; and (2) transmitting this understanding to the individual in a supportive manner and promoting greater satisfaction and acceptance of support in that person" (Pehrson et al., 2016; Kurtz et al., 2017). This scale, based on the content of items 5, 7, and 9, refers to the ability to understand and accept the patient's problems and to transmit this understanding to the patient in a way that leads to a feeling of satisfaction in the patient. It can be related to the challenge component of hardiness. Empathy or the ability to communicate empathetically with patients with AD is an important part of meaningful care and has been shown to enhance the quality of care and health of the caregiver and patient (Brown et al., 2020).

The fourth factor extracted was labeled as "family affective commitment" with 3 items. Family affective commitment refers to the emotional relationship between family members and being responsible to the family (Tice, 2013). Family caregivers, based on their emotional tendencies and having a sense of responsibility toward the family, engage in the process of caring and maintain the caregiver role despite hardships. Therefore, this factor is related to the commitment component of hardiness.

The final extracted factor was labeled "purposeful interaction" with 4 items. Purposeful interaction, based on its definition in the literature (Mehall, 2021), refers to the caregiver's ability to connect with physicians, nurses, family members, and friends to gain information and to improve caregiving abilities and reduce the burden of care and situational stress. According to Maddi's suggestion, the connection can be introduced as the fourth component of the hardiness concept. Maddi believed that interpersonal connection could be an important and influential

TABLE 5 | The results of HTMT of FCHS.

Factors	Religious coping	Self-management	Empathic communication	Family affective commitment	Purposeful interaction
Religious coping					
Self- management	0.755				
Empathic communication	0.526	0.718			
Family affective commitment	0.549	0.643	0.527		
Purposeful interaction	0.429	0.610	0.578	0.296	

TABLE 6 | The results of stability, SEM, responsiveness, and interpretability.

	ICC	SD_{pooled}	Mean	SEM	MDC95	MIC	LOA
Scale	0.903	9.31	86.70	2.89	8.01	4.65	68.45 to 104.94

factor in people's hardiness in dealing with stressful situations because people gain their strength and ability to deal with stressful situations as a result of connecting with others such as family members and members of society. Based on the items' content of this factor, family caregivers of patients with AD also strive to develop their ability to cope effectively with the stresses and challenges of care by communicating with others and gaining information.

LIMITATIONS

One of the important limitations was the concern about the generalization of finding because samples were recruited from Iranian populations. Since culture was recognized as the main factor that affects family caregivers' hardiness, this scale should be tested in other cultures. Therefore, another limitation related to using the online questionnaire for data gathering is that it is not possible to verify the participants' answers due to the lack of physical contact.

STUDY STRENGTH

Nevertheless, this study has several strengths. One of the important strengths is the innovative methodological approach such as Horn's parallel analysis and exploratory graph analysis for extracting factor structure. Furthermore, this study assessed the important and required domain of COSMIN CHECKLIST, namely, the assessment of SEM, ICC, responsiveness, and interpretability that had not been reported previously about hardiness scales.

IMPLICATION

The phenomenon of aging and age-related problems such as AD is increasing, and caring for these patients is an overwhelming and a stressful task for family caregivers. Therefore, being aware of the level of the hardiness of caregivers and designing an intervention to improve hardiness can prevent negative complications and help improve the quality of care. Therefore, the FCHS with the fewer items, good variance explained, and being exclusive for this group is a useful scale for nurses, therapists, and researchers.

CONCLUSION

The finding of this study showed that the FCHS has five dimensions that can be categorized into three components of the Kobasa model including family affective commitment (related to commitment), self-management (related to control), empathic communication (related to challenge), and two new dimensions proposed for this concept including purposeful interaction (related to connection), and religious coping (related to culture). Also, the FCHS scale is a reliable and valid scale with 21 items for assessing the hardiness concept in family caregivers. Based on the results, culture, especially caregivers' beliefs, their ability to manage themselves with patience and positive thinking, communicating with others to raise awareness, and commitment to the family have the most effect on their hardiness.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed in the course of this study are available from the corresponding author on reasonable request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Iran University of Medical Sciences (IR.IUMS.REC.1398.1229). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors contributed in all of stages of this study such as design the study, data collection, analyzing the results, writing the manuscripts, and approving the final manuscript.

ACKNOWLEDGMENTS

We thank all experts and caregivers who contributed to this study.

REFERENCES

Abdollahi, A., and Abu Talib, M. (2015). Hardiness, spirituality, and suicidal ideation among individuals with substance abuse: The moderating role of gender and marital status. *J. Dual Diag.* 11, 12–21. doi: 10.1080/15504263.2014.988558

Abdollahi, A., Hosseinian, S., Zamanshoar, E., Beh-Pajooh, A., and Carlbring, P. (2018). The moderating effect of hardiness on the relationships between problem-solving skills and perceived stress with suicidal ideation in nursing students. *Stud. Psychol.* 60, 30–41.

Alzheimer's Disease International [ADI] (2017). 2017 Alzheimer's disease facts and figures. *Alzheimers Dement.* 13, 325–373. doi: 10.1016/j.jalz.2017.02.001

Armstrong, N. M., Gitlin, L. N., Parisi, J. M., Roth, D. L., and Gross, A. L. (2019). Association of physical functioning of persons with dementia with caregiver burden and depression in dementia caregivers: An integrative data analysis. *Aging Ment. Health* 23, 587–594. doi: 10.1080/13607863.2018.1441263

Ashrafizadeh, H., Gheibizadeh, M., Rassouli, M., Hajibabaee, F., and Rostami, S. (2021). Explain the Experience of Family Caregivers Regarding Care of Alzheimer's Patients: A Qualitative Study. *Front. Psychol.* 12:699959. doi: 10.3389/fpsyg.2021.699959

Baharudin, A. D., Din, N. C., Subramaniam, P., and Razali, R. (2019). The associations between behavioral-psychological symptoms of dementia (BPSD) and coping strategy, burden of care and personality style among low-income caregivers of patients with dementia. *BMC Public Health* 19:447. doi: 10.1186/s12889-019-6868-0

Barlow, J., Wright, C., Sheasby, J., Turner, A., and Hainsworth, J. (2002). Self-management approaches for people with chronic conditions: a review. *Patient Educ. Couns.* 48, 177–187. doi: 10.1016/s0738-3991(02)00032-0

Benishek, L. A., and Lopez, F. G. (2001). Development and initial validation of a measure of academic hardiness. *J. Career Assess.* 9, 333–352.

Brown, E. L., Agronin, M. E., and Stein, J. R. (2020). Interventions to enhance empathy and person-centered care for individuals with dementia: a systematic review. *Res. Gerontol. Nurs.* 13, 158–168. doi: 10.3928/19404921-20191028-01

Cheng, Y.-H., Tsai, C.-C., and Liang, J.-C. (2019). Academic hardiness and academic self-efficacy in graduate studies. *High. Educ. Res. Dev.* 38, 907–921.

Clark, P. (2002). Effects of individual and family hardiness on caregiver depression and fatigue. *Res. Nurs. Health* 25, 37–48. doi: 10.1002/nur.10014

DiBartolo, M. C., and Soeken, K. L. (2003). Appraisal, coping, hardiness, and self-perceived health in community-dwelling spouse caregivers of persons with dementia. *Res. Nurs. Health* 26, 445–458. doi: 10.1002/nur.10107

Ebadi, A., Zarshenas, L., Rakhshan, M., Zareiyan, A., Sharifnia, S., and Mojahedi, M. (2020). *Principles of Scale Development in Health Science*. Tehran: Jame-e-negar.

Elo, S., and Kyngäs, H. (2008). The qualitative content analysis process. *J. Adv. Nurs.* 62, 107–115.

Eschleman, K. J., Bowling, N. A., and Alarcon, G. M. (2010). A meta-analytic examination of hardiness. *Int. J. Stress Manage.* 17, 277–307. doi: 10.1371/journal.pone.0069526

Fujihara, S., Inoue, A., Kubota, K., Yong, K. F. R., and Kondo, K. (2019). Caregiver burden and work productivity among japanese working family caregivers of people with dementia. *Int. J. Behav. Med.* 26, 125–135. doi: 10.1007/s12529-018-9753-9

Henseler, J., Ringle, C. M., and Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *J. Acad. Market. Sci.* 43, 115–135. doi: 10.1007/s11747-014-0403-8

Hooker, K., Monahan, D. J., Bowman, S. R., Frazier, L. D., and Shifren, K. (1998). Personality counts for a lot: Predictors of mental and physical health of spouse caregivers in two disease groups. *J. Gerontol. Ser. Psychol. Sci. Soc. Sci.* 53, 73–85. doi: 10.1093/geronb/53b.2.p73

Hosseini, L., Sharif Nia, H., and Ashghali Farahani, M. (2021). Hardiness in Family Caregivers During Caring From Persons With Alzheimer's Disease: A Deductive Content Analysis Study. *Front. Psychiatry* 12:770717. doi: 10.3389/fpsyt.2021.770717

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equat. Model. Multidiscipl. J.* 6, 1–55.

Kelly, D. R., Matthews, M. D., and Bartone, P. T. (2014). Grit and hardiness as predictors of performance among West Point cadets. *Milit. Psychol.* 26, 327–342.

Kobasa, S. C. (1979). Stressful life events, personality, and health: an inquiry into hardiness. *J. Pers. Soc. Psychol.* 37, 1–11. doi: 10.1037/0022-3514.37.1.1

Koenig, H. G., Pargament, K. I., and Nielsen, J. (1998). Religious coping and health status in medically ill hospitalized older adults. *J. Nerv. Ment. Dis.* 186, 513–521. doi: 10.1097/00005053-199809000-00001

Kurtz, S., Silverman, J., Draper, J., van Dalen, J., and Platt, F. W. (2017). *Teaching and Learning Communication Skills in Medicine*. Landon: CRC press. doi: 10.1201/9781315378398

Lang, A., Goulet, C., and Amsel, R. (2003). Lang and Goulet hardiness scale: Development and testing on bereaved parents following the death of their fetus/infant. *Death Stud.* 27, 851–880. doi: 10.1080/716100345

Lawshe, C. H. (1975). A quantitative approach to content validity. *Pers. Psychol.* 28, 563–575. doi: 10.1097/HMR.0000000000000243

Lincoln, Y. S., and Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New Dir. Program Eval.* 1986, 73–84.

Lynch, S. H., Shuster, G., and Lobo, M. L. (2018). The family caregiver experience–examining the positive and negative aspects of compassion satisfaction and compassion fatigue as caregiving outcomes. *Aging Ment. Health* 22, 1424–1431. doi: 10.1080/13607863.2017.1364344

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (1999). Sample size in factor analysis. *Psychol. Methods* 4, 84–99. doi: 10.1037/1082-989x.4.1.84

Maddi, S. R. (2002). The story of hardiness: Twenty years of theorizing, research, and practice. *Consult. Psychol. J. Pract. Res.* 54:173. doi: 10.1037/1061-4087.54.3.173

Mehall, S. (2021). Purposeful interpersonal interaction and the point of diminishing returns for graduate learners. *Internet High. Educ.* 48:100774. doi: 10.1016/j.iheduc.2020.100774

Moreno-Jiménez, B., Rodríguez-Muñoz, A., Hernández, E. G., and Blanco, L. M. (2014). Development and validation of the Occupational Hardiness Questionnaire. *Psicothema* 26, 207–214. doi: 10.7334/psicothema2013.49

Mund, P. (2017). Hardiness and culture: a study with reference to 3 Cs of Kobasa. *Int. Res. J. Manage. IT Soc. Sci.* 4, 152–159.

Pahlevan, S. S., and Sharif, N. H. (2021). *Factor Analysis and Structural Equation Modeling with SPSS and AMOS. 2*. Tehran: Jamee-negar.

Pashaki, N. J., Mohammadi, F., Jafaraghaee, F., and Mehrdad, N. (2015). Factors influencing the successful aging of Iranian old adult women. *Iran. Red Crescent Med. J.* 17:e22451. doi: 10.5812/ircmj.22451v2

Pehrson, C., Banerjee, S. C., Manna, R., Shen, M. J., Hammonds, S., Coyle, N., et al. (2016). Responding empathically to patients: Development, implementation, and evaluation of a communication skills training module for oncology nurses. *Patient Educ. Couns.* 99, 610–616. doi: 10.1016/j.pec.2015.11.021

Santos da Silva, M. I., de Oliveira Alves, A. N., Barros Leite Salgueiro, C. D., and Bezerra Barbosa, V. F. (2018). Alzheimer's Disease: Biopsycosocial Repercussions in the Life of the Family Caregiver. *J. Nurs. Rev. Enferm.* 12, 1931–1939.

Sharif Nia, H., Kaur, H., Fomani, F. K., Rahmatpour, P., Kaveh, O., Pahlevan Sharif, S., et al. (2021). Psychometric Properties of the Impact of Events Scale-Revised (IES-R) Among General Iranian Population During the COVID-19 Pandemic. *Front. Psychiatry* 12:692498. doi: 10.3389/fpsyt.2021.692498

Sharif Nia, H., Shafipour, V., Allen, K.-A., Heidari, M. R., Yazdani-Charati, J., and Zareiyan, A. (2019). A second-order confirmatory factor analysis of the moral distress scale-revised for nurses. *Nurs. Ethics* 26, 1199–1210. doi: 10.1177/0969733017742962

Sharifi, F., Fakhrzadeh, H., Varmaghani, M., Arzaghi, S. M., Khoei, M. A., Farzadfar, F., et al. (2016). Prevalence of dementia and associated factors among older adults in Iran: National Elderly Health Survey (NEHS). *Arch. Iran. Med.* 19, 838–844.

Soheili, F., Hosseinian, S., and Abdollahi, A. (2021). Development and Initial Validation of the Children's Hardiness Scale. *Psychol. Rep.* 124, 1932–1949. doi: 10.1177/0033294120945175

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties

of health status questionnaires. *J. Clin. Epidemiol.* 60, 34–42. doi: 10.1016/j.jclinepi.2006.03.012

Tho, N. D. (2019). Business students' hardiness and its role in quality of university life, quality of life, and learning performance. *Educ. Train.* 61, 374–386. doi: 10.1108/et-03-2018-0068

Tice, H. (2013). *Development of a Family Affective Commitment Scale*. Omaha, NE: University of Nebraska at Omaha.

Trevisan, K., Cristina-Pereira, R., Silva-Amaral, D., and Aversi-Ferreira, T. A. (2019). Theories of Aging and the Prevalence of Alzheimer's Disease. *BioMed Res. Int.* 2019:9171424. doi: 10.1155/2019/9171424

Weigold, I. K., Weigold, A., Kim, S., Drakeford, N. M., and Dykema, S. A. (2016). Assessment of the psychometric properties of the Revised Academic Hardiness Scale in college student samples. *Psychol. Assess.* 28, 1207–1219. doi: 10.1037/pas0000255

World Health Organization [WHO] (2020). *Ageing*. Geneva: World Health Organization.

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

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 06 April 2022
doi: 10.3389/fpsyg.2022.822931

Measuring Social Desirability in Collectivist Countries: A Psychometric Study in a Representative Sample From Kazakhstan

Kaidar Nurumov[1], Daniel Hernández-Torrano[2], Ali Ait Si Mhamed[2] and Ulzhan Ospanova[1]*

[1] JSC Information-Analytic Center, Nur-Sultan, Kazakhstan, [2] Graduate School of Education, Nazarbayev University, Nur-Sultan, Kazakhstan

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Rodrigo Schames Kreitchmann,
Autonomous University of Madrid, Spain
Juan Carlos Marzo Campos,
Miguel Hernández University of Elche, Spain

***Correspondence:**
Kaidar Nurumov
k.nurumov@gmail.com

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

Received: 26 November 2021
Accepted: 08 March 2022
Published: 06 April 2022

Citation:
Nurumov K, Hernández-Torrano D, Ait Si Mhamed A and Ospanova U (2022) Measuring Social Desirability in Collectivist Countries: A Psychometric Study in a Representative Sample From Kazakhstan. Front. Psychol. 13:822931. doi: 10.3389/fpsyg.2022.822931

Social desirability bias (SDB) is a pervasive measurement challenge in the social sciences and survey research. More clarity is needed to understand the performance of social desirability scales in diverse groups, contexts, and cultures. The present study aims to contribute to the international literature on social desirability measurement by examining the psychometric performance of a short version of the Marlowe-Crowne Social Desirability Scale (MCSDS) in a nationally representative sample of teachers in Kazakhstan. A total of 2,461 Kazakhstani teachers completed the MCSDS – Form C in their language of choice (i.e., Russian or Kazakh). The results failed to support the theoretical unidimensionality of the original scale. Instead, the results of Random Intercept Item Factor Analysis model suggest that the scale answers depend more on the method factor rather than the substantial factor that represents SDB. In addition, an alternative explanation indicates that the scale seems better suited to measuring two SDB correlated factors: attribution and denial. Internal consistency coefficients demonstrated unsatisfactory reliability scores for the two factors. The Kazakhstani version of the MCSDS – Form C was invariant across geographic location (i.e., urban vs. rural), language (i.e., Kazakh vs. Russian), and partially across age groups. However, no measurement invariance was demonstrated for gender. Despite these limitations, the analysis of the Kazakhstani version of the MCSDS – Form C presented in this study constitutes a first step in facilitating further research and measurement of SDB in post-Soviet Kazakhstan and other collectivist countries.

Keywords: social desirability bias, Marlowe-Crowne, MCSDS, validation, Kazakhstan, collectivist culture

INTRODUCTION

Self-reports are an essential tool in the social sciences and the most commonly used assessment and data collection instruments in disciplines such as psychology (Robins et al., 2007), education (Falchikov and Boud, 1989), and sociology (Clair and Wasserman, 2007). The popularity of self-report measures arises from their easy interpretability and administration, the richness of

information, motivation to reflect on the self, and sheer practicality (Paulhus and Vazire, 2007, p. 227). However, the self-report method has been a frequent target of criticism. One of the most vigorous controversies around self-report assessment has been concerning social desirability bias (SDB), or the widespread tendency of individuals to present themselves most favorably with respect to social values and norms (Tracey, 2016).

Social desirability bias has indeed been a concern in personality psychology and survey research since the mid-20th century. Edwards (1957) viewed social desirability as a single dimension that can describe all personality statements. Individuals who obtain high values on the continuum are regarded to have high socially desirable responses. On the contrary, individuals with low values demonstrate low levels of social desirability. From a sociological point of view, "...social desirability as a response determinant refers to the tendency of people to deny socially undesirable traits or qualities and to admit to socially desirable ones" (Phillips and Clancy, 1972, p. 923). Consequently, the presence of socially desirable responses in self-report data is problematic and may lead to spurious correlations between variables and the suppression or the artificial alteration of relationships between constructs of interest (King and Bruner, 2000; van de Mortel, 2008).

Several approaches have been proposed in the literature to prevent or reduce SDB, including forced-choice items, neutral items, randomized response techniques, the introduction of the bogus pipeline, self-administered questionnaires, and the use of proxy subjects. In addition to these, researchers have suggested other methods to detect and measure social desirability effects (Nederhof, 1985). Among them, the use of social desirability scales is the most common. Social desirability scales are included in conjunction with the targeted questionnaire(s) as indicators of discriminant validity. Ideally, the correlation between the scores of the targeted questionnaire and the social desirability measure is zero to weak, demonstrating that the variable of interest is unconfounded with social desirability (Tracey, 2016).

Multiple social desirability scales have been developed in past decades (see Paulhus, 1991). The Marlowe-Crowne Social Desirability Scale (MCSDS) (Crowne and Marlowe, 1960) is one of the most widespread scales to measure SDB around the world (Beretvas et al., 2002). It measures social desirability as "the need to obtain approval by responding in a culturally appropriate and acceptable manner" (Crowne and Marlowe, 1960, p. 353). The MCSDS consists of 33 binary items with true or false answers on culturally sanctioned and approved but improbable behaviors (e.g., I have never deliberately said something that hurt someone's feelings). According to Crowne and Marlowe (1964), a unidimensional construct underlies the MCSDS: "need for approval." Thus, higher scores in the MCSDS reflect higher needs for social approval and a tendency to portray yourself more positively.

The psychometric properties of the MCSDS have been widely studied in multiple contexts and cultures, predominantly in North America (Fischer and Fick, 1993; Loo and Thorpe, 2000; Barger, 2002; Loo and Loewen, 2004; Leite and Beretvas, 2005; Ventimiglia and MacDonald, 2012), although studies involving European (Sârbescu et al., 2012; Vésteinsdóttir et al., 2015) and Asian samples (e.g., Seol, 2007) are also available. The factor structure of the scale has been extensively analyzed through exploratory and confirmatory factor analysis, and a few studies have begun to implement alternative approaches such as item response theory and Rasch measurement (Seol, 2007; Vésteinsdóttir et al., 2017). Collectively, these studies provide inconclusive evidence on the dimensionality of the MCSDS. Some studies support the theoretical unidimensionality of the scale (e.g., Seol, 2007; Vésteinsdóttir et al., 2015), while other studies provide stronger evidence for a two-factor structure (e.g., Loo and Loewen, 2004; Ventimiglia and MacDonald, 2012) or alternative factorial solutions (e.g., Loo and Thorpe, 2000; Barger, 2002; Leite and Beretvas, 2005). Reliability analyses have also shown mixed results on the internal consistency of the scores, with coefficients ranging from 0.72 (Loo and Thorpe, 2000) to 0.96 (Fischer and Fick, 1993).

Several short versions of the MCSDS have been developed to avoid excessive item redundancy and length of the full scale (e.g., Strahan and Gerbasi, 1972; Reynolds, 1982; Ballard, 1992). These forms range between 10 and 20 items and result from factor analysis techniques assuming that the MCSDS full version assesses one single dimension. Internal consistency scores of the short versions are lower but comparable to those of the full version. Moreover, they have been considered suitable substitutions and, in some cases, significant improvements in fit over the full scale (Loo and Thorpe, 2000; Barger, 2002; Loo and Loewen, 2004; Sârbescu et al., 2012). The MCSDS – Form C developed by Reynolds (1982) stands out as one of the most commonly used short forms available. It comprises 13 items and demonstrates good psychometric characteristics compared to other short versions. The MCSDS – Form C internal consistency estimates range from 0.62 to 0.89 and its scores correlate strongly with the scores on the full scale (r = 0.91 to 0.96) (Reynolds, 1982; Ballard, 1992; Fischer and Fick, 1993; Loo and Thorpe, 2000; Barger, 2002; Loo and Loewen, 2004; Vésteinsdóttir et al., 2015). However, confirmatory factor analyses have provided conflicting results about the factorial structure of the MCSDS – Form C, with only partial support for the unidimensionality assumption (Barger, 2002; Loo and Loewen, 2004; Leite and Beretvas, 2005; Verardi et al., 2009; Vésteinsdóttir et al., 2015).

The measurement invariance of different versions of the MCSDS has been partially supported in previous studies. For example, Kurz et al. (2016) confirmed measurement invariance between genders in the context of Malaysia. However, the authors found only partial support for measurement invariance across languages in the Chinese and English versions of the MCSDS. Concern has also been raised about the cross-cultural validity of the MCSDS scales. Differences in the tendency to respond in a socially desirable manner across countries and cultural groups have been reported in several studies (e.g., Verardi et al., 2009; He et al., 2015). For example, Middleton and Jones (2000) used the full MCSDS scale in a convenience sample of Western and Eastern university students and found that Eastern participants were more likely to deny socially undesirable traits and to admit socially desirable traits compared to Western participants. Lalwani et al. (2006) tested the hypothesis that collectivist cultures tend to engage in deception and socially

desirable responses more than individualistic cultures. Their findings suggested that people from both types of cultures engage in desirable responses, although in different ways. Individualism seemed to be more associated with the tendency to report inflated views of one's skills and capabilities, while collectivism was linked to the tendency to present self-reported actions in the most positive manner.

More clarity is needed to understand the performance of social desirability scales in diverse groups, contexts, and cultures. The present study aims to contribute to the international literature on the measurement of social desirability by examining the psychometric performance of the MCSDS – Form C in a nationally representative sample of teachers in Kazakhstan. Kazakhstan provides an interesting context to explore social desirability measurement for several reasons. First, the country occupies a strategic geopolitical location in the Eurasian mass and constitutes a unique blend of Eastern and Western cultures. Kazakhstan is in fact a diverse country with more than 120 ethnic groups that have different social values and norms (The Agency on Statistics of the Republic of Kazakhstan, 2011). Second, as a former Soviet republic, Kazakhstan maintains a strong national collectivist tradition (Winter et al., 2020). This is relevant as collectivist cultures tend to demonstrate stronger and more consistent magnitudes and patterns of SDB (Bernardi, 2006; Kim and Kim, 2016). Third, measuring SDB is particularly important in societies that have experienced authoritarian regimes in the past, such as Kazakhstan. Finally, SDB is a widespread problem that affects many areas, including education. Social desirability may explain the questionable results of the latest international evaluations such as TALIS-2018 in the context of Kazakhstan, in which teachers report values well above the OECD average in some questions. For example, 82% of Kazakhstani teachers were confident in their ability to teach using ICT (OECD average of the OECD was 67%). At the same time, 30% of teachers marked ICT for teaching as the main priority of professional development (Information-Analytic Center [IAC], 2019; OECD, 2019). Having a reliable and valid tool to measure SDB could help to account for the measurement error caused by this phenomenon in Kazakhstan, Central Asia, and other collectivistic countries.

MATERIALS AND METHODS

Description of Sample

The sample consisted of subject teachers who participated in the UNESCO Teachers' Readiness Survey in early 2021 in Kazakhstan (Information-Analytic Center [IAC], 2021). The survey is based on the UNESCO ICT competency framework for teachers and covers areas such as teacher ICT competencies, use of ICT in teaching, awareness of the official policy on ICT use in education and professional learning (UNESCO, 2011). To ensure large-scale representativeness, the sample design consisted of an explicit stratified selection of a proportionally allocated sample from the population list of subject teachers, as well as a weighting strategy. The latter included adjustment for unknown eligibility, adjustment for non-response, post-stratification, and extreme weights trimming. In total, 2,851 subject teachers were selected for the main study with a final response rate of 86% (n = 2,461). The weighted sample mean age of subject teachers is 40.58 (std. error = 0.22) whereas the population mean age is 40.50. Additional information on the distribution of the raw sample responses in biographic and geographic subgroups is presented in **Table 1**.

One can notice significant gender disproportion among men – 470 (19%) and women – 1991 (81%). This disproportion is expected due to the traditional overrepresentation of women in school teaching in the context of Kazakhstan. Additionally, the distributions of responses show higher proportions of Kazakh language and rural subject teachers in terms of subgroups of language and geographic location.

Instruments

The Marlowe Crowne Social Desirability Scale (MCSDS) – Form C (Reynolds, 1982) was used to measure social desirability bias in this study. The MCSDS – Form C is a brief questionnaire comprising 13 items that represent a selection of socially desirable and undesirable behaviors (e.g., "No matter who I'm talking to, I'm always a good listener," "There have been occasions when I took advantage of someone"). Items are dichotomously scored on a true/false scale. A score of 1 is granted if the participant responds "true" to a socially desirable item or "false" to a socially undesirable item. On the contrary, a score of 0 is provided if the participant responds "false" to a socially desirable item or "true" to a socially undesirable item. A total score can be obtained summing up the scores for all items, with higher scores representing higher SDB.

The MCSDS – Form C was translated into the two official languages of Kazakhstan (i.e., Russian and Kazakh) using a back-translation approach (Brislin, 1970). In addition to that, the Russian and Kazakh translations of the MCSDS – Form C were further assessed by the research team to ensure understandability, psychological equivalence, and the accuracy

TABLE 1 | Distribution of raw sample responses in subgroups.

	n	%
Gender		
Male	470	19.0
Female	1,991	81.0
Language		
Kazakh	1,507	61.2
Russian	954	38.8
Geographic locality		
Rural	1,422	57.8
Urban	1,039	42.2
Age groups		
18–35 years	914	38.0
36–50 years	997	41.4
51–72 years	496	20.6

Age was transformed into the categorical variable with three categories. A 51–72 years old group though smallest in terms of the number of teachers, nonetheless, includes a larger range in years than 18–35 and 36–50 groups. This is due to a skewed population distribution toward younger teachers.

of the translations. The MCSDS – Form C was included in the UNESCO questionnaire and distributed online. Anonymity and confidentiality were ensured, no information that could identify the identities of the participants was collected.

Procedure and Data Analysis

Descriptive analyses were used to describe the pattern of responses on the MCSDS – Form C. In addition, the tetrachoric correlation matrix between the items was calculated. Tetrachoric correlation is a special case of polychoric correlation specifically used with ordinal dichotomous data (Pearson, 1900; Carrol, 1961), as is the case in the MCSDS – Form C. Furthermore, to test the psychometric performance of the MCSDS – Form C in Kazakhstan, we used a five-step approach that included (1) dimensionality reduction, (2) exploration of factorial structure, (3) confirmation of factorial structure, (4) analysis of measurement invariance across gender, age, language, and geographic location, and (5) factorial and composite reliability analysis (see **Figure 1**).

The factorial structure of the MCSDS – Form C was first examined using several dimensionality reduction approaches. First, a Principal Component Analysis (PCA) was implemented on the matrix of tetrachoric correlations. The Kaiser criterion, the results of parallel analyses, and the interpretation of the scree plot were used to determine the number of factors underlying the structure of the scale. Second, a Categorical Principal Component Analysis (CATPCA) conducted on the raw data was used to further explore the dimensionality of the scale. CATPCA is a technique of optimal scaling designed specifically for categorical ordinal and nominal data with the ability to account for non-linear relations between variables. Instead of a linear combination of transformed variables, the method transforms, through iterative computation, the matrix of actual categorical data into quantified data with further maximization of eigenvalues on the matrix of quantified data (Gifi, 1990; Linting et al., 2007).

The resulting dimensions were further analyzed using an Exploratory Factor Analysis (EFA) computed on the matrix of tetrachoric correlations. The robust weighted least squares (WLS) estimator was used to account for the dichotomous nature of the scale. The robust version uses only diagonal elements of the weight matrix to obtain standard errors (Muthen et al., 1997), whereas the standard version employs a full weight matrix (Browne, 1984). Both robust and standard estimators are asymptotically free. However, the robust WLS shows stable results in samples of different sizes, while the standard WLS shows stability only in large samples (Flora and Curran, 2004; Barendese et al., 2014).

The resulting factor structures were tested using a Confirmatory Factor Analyses (CFA) correlated factor models with a diagonally weighted least square estimator (DWLS), as suggested by Brown (2006). In addition, we tested alternative, more complex factor structures such as bifactor and hierarchical factor models. The former allows to model separate effects of specific and general factors while the later accounts for the direct effect of the higher order factor on the first order factors. The Chi-square test (χ^2) was used to evaluate the absolute fit of the model. However, because the χ^2 test is considered highly conservative, additional fit indices were used to evaluate the model, such as the Comparative Fit Index (CFI), the Tucker-Lewis Index (TLI), and the Root Mean Square Error of Approximation (RMSEA). The values of CFI and TLI > 0.95 and RMSEA < 0.06 indicated a good model fit, while CFI and TLI > 0.90 and RMSEA < 0.08 indicated a satisfactory fit (Hu and Bentler, 1999; Schreiber et al., 2006). Finally, to offer an alternative account of the factorial structure of the scale, we conducted a Random Intercept Item Factor Analysis (RIIFA) to test whether the results of the MCSDS – Form C contain a method factor along with the substantial factor representing social desirability. For instance, this can be due to negatively and positively worded items (Marsh, 1996; DiStefano and Motl, 2009) in survey instruments. The effect of a method factor can be found via modeling residual covariance separately between positive and negative items (Marsh, 1989, 1996) or by allowing intercept in a CFA model to vary across respondents in a Random Intercept Item Factor Analysis (RIIFA, Maydeu-Olivares and Coffman, 2006; Nieto et al., 2021). In the latter, one needs to add one method factor and set its loadings to 1 with free estimated variance. The approach is appropriate to model individual styles of responses and helps to identify whether a multidimensional structure is truly due to substantive factors or due to a spurious, method factor which goes along with the substantive factor. Hence, we run an additional RIIFA model and check the fit statistic and variance of the random component.

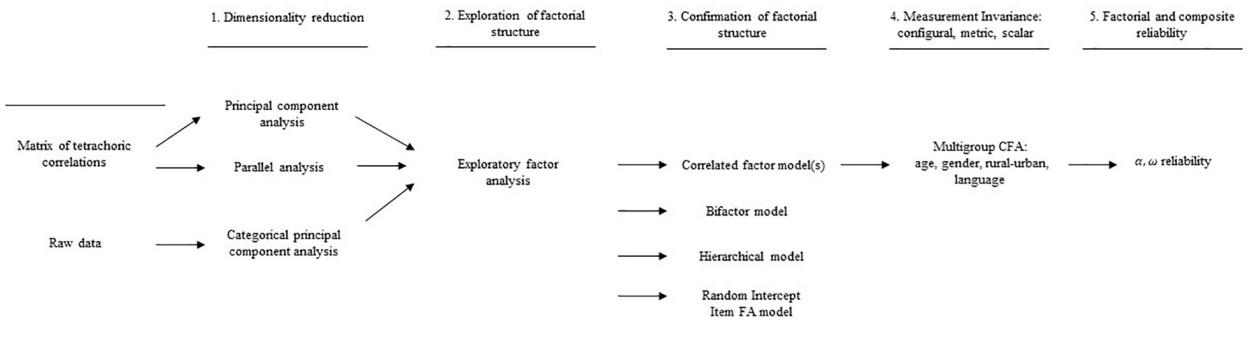

FIGURE 1 | Psychometric properties of MCSDS – Form C analysis flowchart.

Further, we tested configural (unconstrained), metric (constrained slopes), and scalar (constrained slopes and intercepts) measurement invariance across gender, age, language, and geographic location using Multiple Group Confirmatory Factor Analysis (MGCFA). The likelihood ratio test was used to compare statistically significant changes between different models at the $p < 0.05$ level. A non-statistically significant change was interpreted as the indication supporting measurement invariance (Satorra and Bentler, 2000).

Finally, after exploring the dimensionality and testing the measurement invariance of the scale, we examined the factorial and composite reliability of the scores. To investigate the reliability of the Kazakhstani version of the MCSDS – Form C, we calculated the Cronbach alpha coefficient on the matrix of tetrachoric correlations of the full scale. However, when the instrument does not have Tau-Equivalent items (equal factor loadings) and shows multidimensionality, alpha is not the optimal solution. Moreover, the alpha coefficient often serves as a lower bound or largely underestimates reliability (Sijtsma, 2009). Furthermore, when multidimensionality is detected via the CFA framework, a more appropriate alternative is to use the omega reliability coefficient (McDonald, 1999; Green and Yang, 2015; Flora, 2020). Omega calculates reliability of the scale that is due to the presence of some general factor in bifactor and hierarchical models as well as group-specific factors (Green and Yang, 2015). In this study, we focus on composite reliability, or in other words, the sum of factor loadings of individual items. We calculated the ω coefficient for correlated factors in the CFA models and also show the composite alpha coefficient.

All calculations were carried out with the R statistical programming language (R Core Team, 2020). The PCA was performed using the *FactoMiner* package with PCA function (Le et al., 2008). The CATPCA was performed using the *gifi* package and the *princals* function (Mair et al., 2019). EFAs were performed using the *psych* package, with the *fa* function (Revelle, 2021). CFA and measurement invariance tests were calculated using the specialized package for structural equation modeling *lavaan* (Version 0.6-9; Rosseel, 2012). Reliability analysis was calculated with the *SEMTools* package (Version 0.5-5; Jorgensen et al., 2021). The R scripts with all calculations are provided as **Supplementary Material**.

RESULTS

Descriptive Statistics

The response pattern for the MCSDS – Form C items is presented in **Table 2**. We recalculated socially desirable responses as 1 (socially desirable response is detected) and 0 (no socially desirable response is detected). In the table, the dichotomy is presented in the form of "yes" and "no." In general, the results suggest high levels of social desirability bias for all items, except items 1 (59.6%) and 2 (49.0%). **Table 2** also depicts the matrix of tetrachoric correlations between the items. The correlation ranges from low negative $r_{tet} > -0.1$ between items 13 and 12 to moderate positive $r_{tet} < 0.58$ between items 7 and 5. For some pairs of items (e.g., 13 and 2, 13, and 3), the correlation is essentially 0, suggesting the absence of statistical interdependence.

Dimensionality Reduction

Table 3 shows the PCA results on the matrix of tetrachoric correlations for the first five components. The analysis yielded three components with eigenvalues greater than 1, accounting for 72.33% of the total variance. However, the leveling of the eigenvalues on the scree plot and the results of the parallel analysis do not provide a definitive answer to the dimensionality of the scale (see **Figure 2**).

Both the two- and the three-component solutions appear as plausible solutions. Alternatively, we explored the dimensionality of the scale by running CATPCA on the actual data. As in linear PCA, we looked at eigenvalues and the explained variance or variance accounted for (VAF) to understand how many components to retain. Furthermore, eigenvalues larger than 1, as well as the scree plot, can help to decide the adequate number of components (Linting et al., 2007). The results suggest at least two clear dimensions with eigenvalues of 2.27 and 1.67 and a cumulative variance explained of 30.33%. With the inclusion of the third component with an eigenvalue of 1.12, the cumulative variance increases from 30.33 to 38.95%. **Figure 3** also suggests at least two clear components with a plausible additional third component. Overall, the results of the dimensionality reduction techniques suggest the existence of two or three components underlying the structure of the MCSDS – Form C.

Exploration of Factorial Structure

The two- and three-component structures were further examined using EFAs with oblique rotation on the matrices of tetrachoric correlations. The results of the EFA for the two- and three-factorial solutions are presented in **Table 4**. The two-factor solution demonstrated acceptable loadings (i.e., >0.40) for the 13 items of the MCSDS – Form C. Eight items load on factor 1, which explained 20% of the variance. Five items demonstrated loadings on factor 2, accounting for 17% of the variance. The high uniqueness of item 13 is noteworthy (0.84). In addition, item 6 and item 8 load on both factors, although loadings on factor 1 are at least two times larger than on factor 2. The correlation between the two factors was modest ($r = 0.22$).

The three-factorial solution achieved similarly acceptable item loadings. The same eight items loaded into factor 1. The remaining items loaded into factor 2 (3 items) and factor 3 (2 items). Factors 1, 2, and 3 explained 20, 15, and 7% of the total variance, respectively. Since we allowed factors to correlate, one can notice that items 6, 8, and 10 have additional loadings on factor 2. There was a moderate correlation between factor 1 and factor 2 ($r = 0.27$) and between factor 2 and factor 3 ($r = 0.26$). However, no statistically significant relationship was found between factor 1 and factor 3 ($r = 0.02$).

Overall, the results of the EFAs suggest that these factorial structures could be a result of theoretical dimensions of SDB but also due to methodological influences related to the keyed direction of the items of the scale. In the next section, several factor theoretical and methodological solutions are tested using CFAs.

TABLE 2 | Pattern of responses across items and matric of tetrachoric correlation (n = 2,407).

	Yes (%)	No (%)	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12
Item 1	59.6	40.4	–											
Item 2	49.0	51.0	0.50	–										
Item 3	65.8	34.2	0.35	0.42	–									
Item 4	85.6	14.4	0.37	0.41	0.39	–								
Item 5	94.1	5.9	−0.03	−0.03	−0.06	−0.01	–							
Item 6	91.3	8.7	0.30	0.26	0.18	0.37	0.20	–						
Item 7	96.8	3.2	0.03	−0.03	−0.09	0.01	0.58	0.23	–					
Item 8	89.2	10.8	0.14	0.24	0.27	0.33	0.14	0.45	0.17	–				
Item 9	90.6	9.4	0.06	0.12	0.07	0.05	0.50	0.16	0.53	0.26				
Item 10	70.8	29.2	0.05	0.15	0.08	0.06	0.33	0.11	0.40	0.12	0.36	–		
Item 11	84.7	15.3	0.24	0.34	0.33	0.28	0.10	0.40	0.07	0.32	0.06	0.14	–	
Item 12	77.5	22.5	0.30	0.34	0.25	0.34	0.15	0.35	0.17	0.33	0.17	0.11	0.26	–
Item 13	72.8	27.2	−0.01	0.00	0.00	0.03	0.19	0.13	0.26	0.10	0.27	0.45	−0.07	−0.09

TABLE 3 | Results of PCA and CATPCA (n = 2.407).

	Linear PCA			CATPCA		
Component	**Eigenvalue**	**% of variance explained**	**Cumulative% of variance explained**	**Eigenvalue**	**% of variance explained**	**Cumulative% of variance explained**
1	6.541	50.315	50.315	2.27	17.50	17.50
2	1.807	13.904	64.219	1.67	12.82	30.33
3	1.054	8.114	72.334	1.12	8.61	38.95
4	0.757	5.829	78.163	1.03	7.94	46.89
5	0.724	5.576	83.740	0.89	6.85	53.75

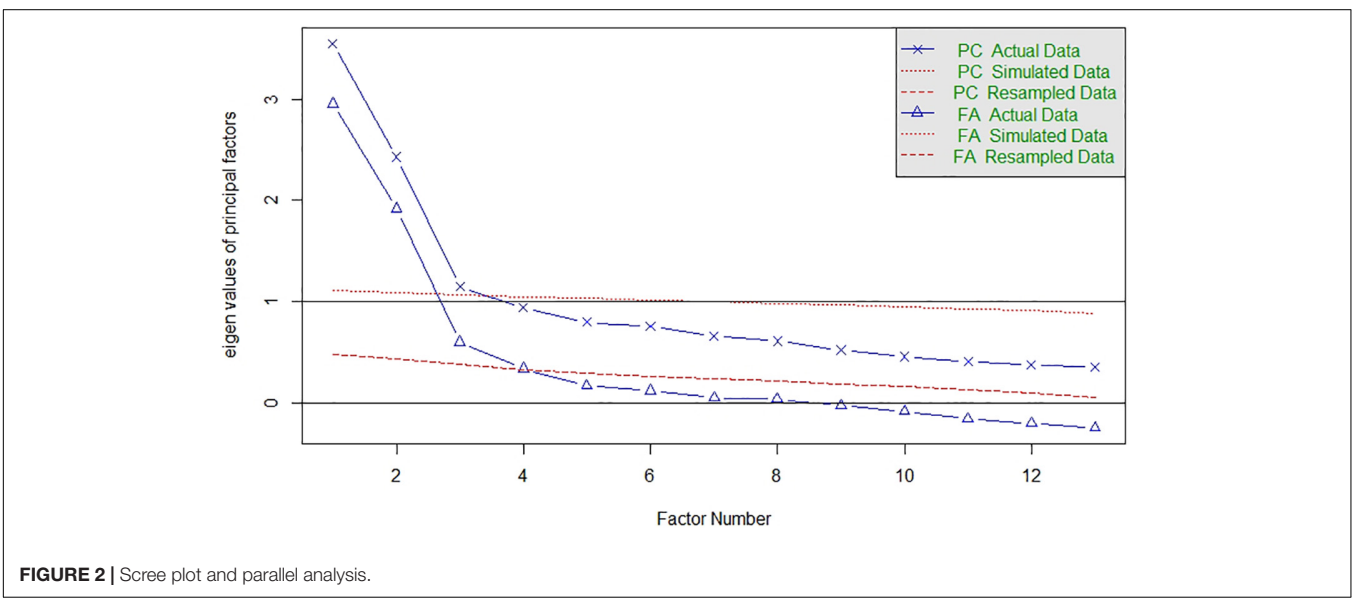

FIGURE 2 | Scree plot and parallel analysis.

Confirmation of Factorial Structure

Confirmatory Factor Analyses were conducted to examine the structural validity of the two-factor and three-factor solutions emerging from the EFA, as well as their more complex alternatives (i.e., bifactor and hierarchical factor models). Furthermore, for reasons of comparison and to test the hypothetical one-factor structure of the MCSDS – Form C, we run a CFA for the unidimensional model. As in the EFA analysis, the parameter estimates in the models were obtained using the robust diagonally weighted least squares (DWLS) estimator to account for the dichotomous nature of MCSDS – Form C. **Table 5** presents the robust fit indices of the calculated models. As indicated by the χ^2 values, none of the models fit perfectly. In line with the multidimensional structure revealed in previous

TABLE 4 | Results of the EFAs for the two- and three-factorial solutions (n = 2,407).

	Two-factor model				Three-factor model				
	Factor 1	Factor 2	h2	μ2	Factor 1	Factor 2	Factor 3	h2	μ2
Item 1	0.60		0.34	0.66	0.61			0.35	0.65
Item 2	0.68		0.45	0.55	0.71	−0.20		0.48	0.52
Item 3	0.60		0.34	0.66	0.62	−0.22		0.36	0.64
Item 4	0.64		0.40	0.60	0.64			0.40	0.60
Item 5		0.72	0.50	0.50		0.77		0.56	0.44
Item 6	0.52		0.35	0.65	0.49	0.25		0.37	0.63
Item 7		0.79	0.61	0.39		0.76		0.62	0.38
Item 8	0.47		0.29	0.71	0.45	0.22		0.30	0.70
Item 9		0.67	0.47	0.53		0.58	0.20	0.45	0.55
Item 10		0.52	0.29	0.71			0.55	0.47	0.53
Item 11	0.53		0.28	0.72	0.51	0.20		0.29	0.71
Item 12	0.52		0.30	0.70	0.50			0.35	0.65
Item 13		0.41	0.16	0.84			0.66	0.48	0.52

Factor loadings < 0.20 are omitted.

TABLE 5 | CFA and RIIFA comparison of standard fit statistics (robust is given in parenthesis, n = 2,407).

Model	RMSEA	TLI	CFI	χ^2	degrees of freedom	*p*-value
One-factor model	0.071 (0.073)	0.709 (0.624)	0.757 (0.686)	856 (887)	65	0
Two-factor model	0.035 (0.036)	0.931 (0.905)	0.943 (0.922)	249 (268)	64	0
Three-factor model	0.030 (0.033)	0.947 (0.922)	0.958 (0.938)	198 (224)	62	0
Bifactor model with two specific factors	0.024 (0.029)	0.967 (0.940)	0.978 (0.959)	126 (160)	53	0
Hierarchical model with three first order factors	0.030 (0.033)	0.947 (0.922)	0.958 (0.938)	198 (224)	62	0
Random Intercept One Factor model	0.032 (0.035)	0.940 (0.914)	0.951 (0.929)	225 (248)	64	0

The bifactor model with three specific factors was tested but failed to be identified. Hierarchical models with two first order factors tend to be underidentified (Brown, 2006) and was therefore not tested in this study.

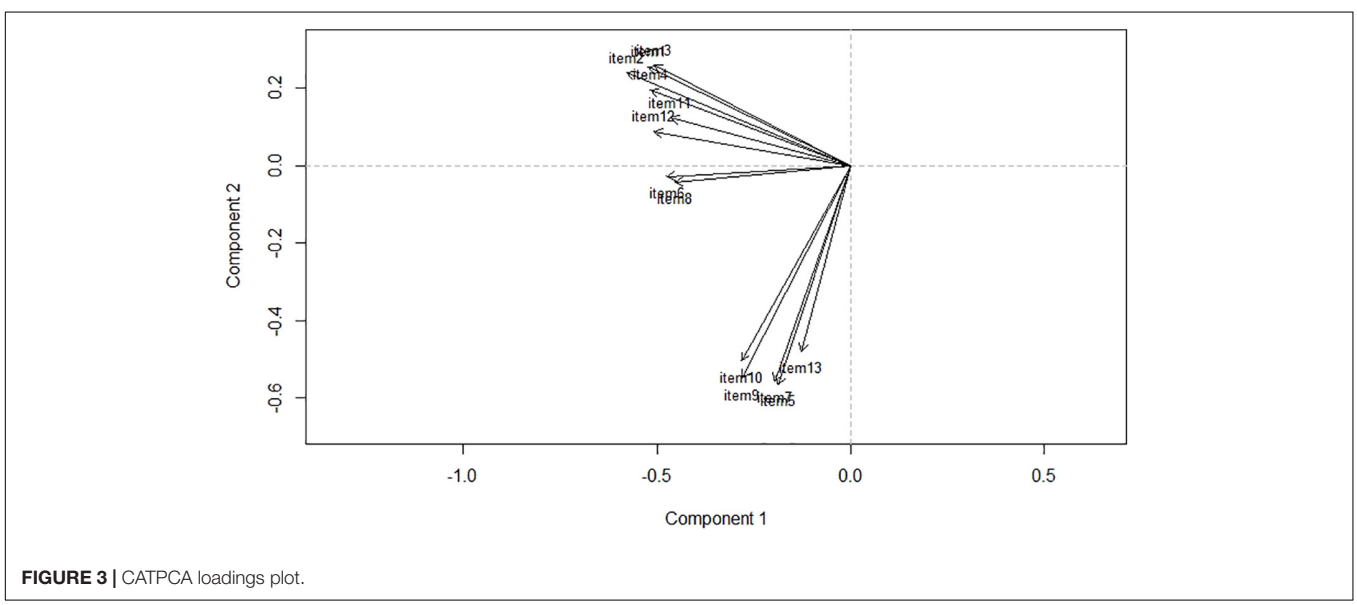

FIGURE 3 | CATPCA loadings plot.

analyses, the unidimensional solution indicated the worst fit. The two-factor model was found to have an absolute satisfactory fit, with standard CFI = 0.94, TLI = 0.93, and RMSEA = 0.035. The three-factor model also achieved a satisfactory fit, with CFI = 0.95, TLI = 0.94, and RMSEA = 0.030. Although both models demonstrated a satisfactory fit, the differences in TLI, CFI, and RMSEA between the two models demonstrated the superiority of the three-factor model. In addition, since we used

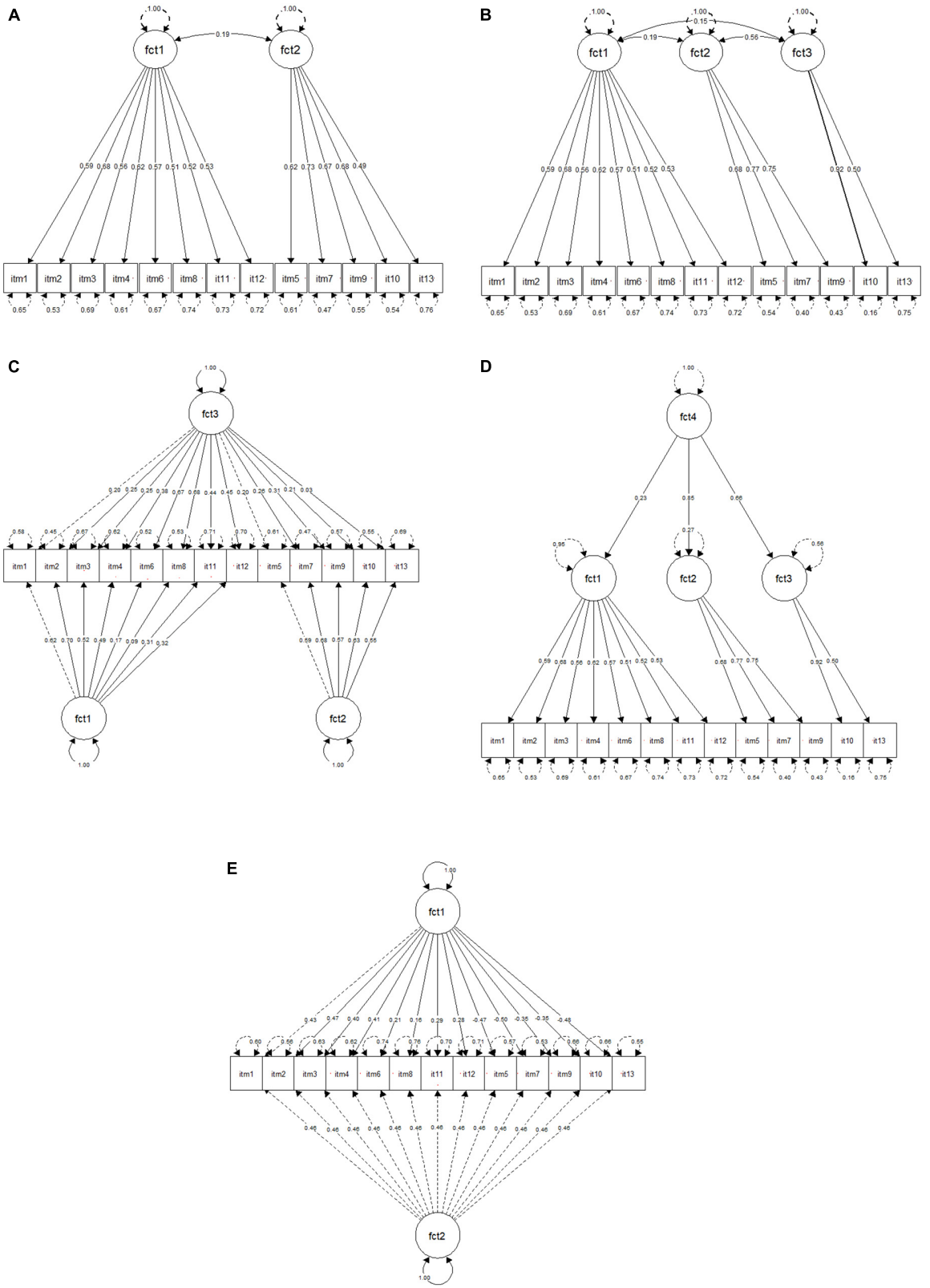

FIGURE 4 | Standardized factor loadings for the two-, three-, bifactor, hierarchical, and random item intercept models of the MCSDS – form C (n = 2,407). **(A)** Two-factor model. **(B)** Three-factor model. **(C)** Bifactor model *. **(D)** Hierarchical three-factor model. **(E)** Random item intercept factor model. * Loading between factor 2 and item 5 is fixed to 1 for model identification. Covariances between specific factors and between general and specific factors are fixed to 0.

the DWLS estimator, the difference between the nested models was calculated with a scaled Satorra-Bentler chi-square difference test (Satorra and Bentler, 2000). In support of the comparison between the fit indices, there was a statistically significant difference between the two- and the three-factor models with a *p*-value of 4.958e-07. **Figures 4A,B** presents the standardized path estimates for both models. All standardized path estimates were significantly loaded into the hypothesized specific factors in the two-factor ($\beta = 0.49$ to 0.73, $p < 0.01$) and three-factor models ($\beta = 0.50$ to 0.77, $p < 0.01$).

The bifactor solution with two specific factors showed the highest TLI = 0.967 and CFI = 0.978 and the lowest RMSEA = 0.024 which indicated the best absolute fit among the calculated models. However, notwithstanding the fit indices, the model had poor loadings (<0.40) between general factor and a set of items, ranging from $\beta = 0.03$ to 0.38 (**Figure 4C**). The bifactor solution with three specific factors failed to be identified. Thus, despite the best absolute fit, the three-factor model can be still regarded as superior to the bifactor solution. We also calculated a hierarchical model with three first order factors and one second order factor. The standard fit statistics of the higher order model produced identical results to the three-factor correlated model. However, it is useful to look at factor estimates as well as loadings between the first and the second order factors (**Figure 4D**). The results of the standardized solution showed weak loading of higher level with factor 1 ($\beta = 0.23$), high but not statistically significant loading with factor 2 ($\beta = 0.85$, *z*-value = 1.801), and moderate high with factor 3 ($\beta = 0.66$). The rest of the loadings between the second order latent variables and items were essentially identical to the three-factor model. Finally, since hierarchical models with two second order factors are considered to be underidentified (Brown, 2006), we did not try to extract the general factor from the two-factor model.

Models examined above accounted for substantive, theory driven factors. **Table 5** presents the fit statistics of the RIIFA model to test the specific variance associated with the item keying as a result of a methodological artifact. Standard fit statistics demonstrated a satisfactory fit for the RIIFA model, with TLI = 0.940, CFI = 0.951, RMSEA = 0.032. In addition, the estimate of random component variance accounted for about 21% of all variances with significant $z = 23.65$ and std. error = 0.009. This is larger than the variance the substantive factor where the estimate is 0.18 with $z = 6.90$ and std. error = 0.027. However, some factor loadings in the RIIFA model were relatively small ($\beta < 0.40$) (see **Figure 4E**).

Based on the findings above, we proceeded to explore measurement invariance for the two and three-factor models. The one-factor model was not further analyzed because of the unsatisfactory fit. Due to low loadings and no statistical significance between latent variables (general and specific) and some observed variables in the bifactor solution as well as first and second order factors in the hierarchical solution, these models were not further tested for measurement invariance and reliability either. Also, the RIIFA was not further explored for measurement invariance and reliability due to the low factor loadings of some of the items (e.g., items 6, 8, 11, and 12).

Measurement Invariance Across Gender, Age, Language, and Geographic Location

The MGCFA results for measurement invariance for the two- and three-factor solution of the MCSDS – Form C across gender (male vs. female), age (18–35 year old vs. 36–50 year old vs. 51–72 year old), language (Russian vs. Kazakh), and geographic location (urban vs. rural) are presented in **Table 6**.

For the two-factor solution, the MGCFA did not show statistical significance and therefore full configural-metric and full metric-scalar invariance for rural and urban teachers with $p = 0.51$ and $p = 0.43$, respectively. The analysis established partial configural-metric invariance ($p = 0.08$) with item 3 being freed up in the constrained loadings model for factor 1 and partial metric-scalar invariance ($p = 0.11$) among teachers from different age groups where in addition to item 3, we allowed loadings between factor 1 and items 4 and 6 to vary between groups. Furthermore, while the analysis did not show statistical significance between the configural and metric models for gender with $p = 0.16$, the invariance between the metric and configural models was not reached ($p = 0.02$). The Lagrange Multiplier Test did not indicate significant items with all *p*-values above the threshold of 0.05. As in the MGCFA analysis for the three-factor solution, the likelihood ratio test between the configural and scalar models did not show statistically significant differences with $p = 0.52$. Finally, for the language group, we found no difference ($p = 0.46$) between the general model with varied intercepts and loadings across Russian and Kazakh speaking teachers and partial invariance ($p = 0.70$) with items 2 and 3 being freed up for factor 1 in the scalar model.

For the three-factor solution, measurement invariance was established between rural and urban participants with $p = 0.50$ between configural – metric and $p = 0.75$ between metric – scalar. The same was true in the Russian-Kazakh language of the questionnaire, with $p = 0.56$ between configural – metric and $p = 0.53$ between metric – scalar. It is important to point out that the scalar model for language showed a statistically significant difference with the metric model and thereby we switched to partial solution freeing up loadings for items 2, 3, and 6 in factor 1. For age, the configural and scalar models failed to demonstrate measurement invariance, as some estimated variances showed negative signs. For gender, we encountered the same problem with metric invariance. However, comparing the configural model with the scalar model, the *p*-value was 0.56.

Overall, these findings demonstrate the measurement invariance of the MCSDS – Form C across language and geographic location for both models, but not across gender groups in the two- and three-factor solutions and age in the three-factor solution.

Factorial and Composite Reliability

Two approaches were implemented to explore the reliability of the scores in the two models under examination for the Kazakhstani version of the MCSDS – Form C. First, internal consistency was examined using Cronbach's alpha (α) coefficient. The results demonstrated adequate internal reliability for the two

TABLE 6 | Measurement invariance.

	Two-factor model		Three-factor model	
Group	**MI**	***p* value**	**MI**	***p* value**
Rural-urban	Configural – metric	0.51	Configural – metric	0.50
	Metric – scalar	0.43	Metric – scalar	0.74
Age	Configural – metric (partial)	0.08	Configural (failed) – metric	–
	Metric (partial) – scalar	0.11	Metric – scalar (failed)	–
Gender	Configural – metric	0.16	Configural – metric (failed)	–
	Metric – scalar (failed)Configural – scalar	–0.52	Metric (failed) – scalarConfigural – scalar	–0.56
Language	Configural – metric	0.46	Configural – metric	0.56
	Metric – scalar (partial)	0.70	Metric – scalar (partial)	0.53

dimensions of the two-factor model (α = 79, α = 76, respectively). For the three-factor model, internal reliability was adequate for factor 1 (α = 79) and factor 2 (α = 77), but lower for factor 3 (α = 62). Second, to account for the multidimensionality of the scale, the reliability of the scores was examined using the McDonald's omega (ω) statistic. Coefficient ω for subscale internal consistency exhibited poor reliability indices for the two dimensions in the two-factor (ω = 0.54, ω = 0.50, respectively). Similarly, coefficient ω for the three dimensions in the three-factor model were low, ranging from 0.47 to 0.54. We do not specifically discuss an acceptable threshold of reliability in this paper, but we expect group-specific factors to be higher than 0.70 to be counted as at least acceptable.

DISCUSSION

This research investigated the psychometric performance of the Marlowe-Crowne Social Desirability Scale (MCSDS) – Form C in a nationally representative sample of teachers in Kazakhstan. We examined the factorial structure of the scale using several dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Categorical Principal Component Analysis (CATPCA), as well as Exploratory Factor Analysis (EFA) computed on the matrix of tetrachoric correlations. Furthermore, the theoretical structure of the scale was further tested using a Confirmatory Factor Analysis (CFA) and a Random Intercept Item Factor Analysis (RIIFA). We tested whether the measure varied between gender, age, geographic location, and language groups using Multigroup Confirmatory Factor Analyses (MGCFA). Finally, the reliability of the scores was explored using Cronbach's alpha and McDonald's omega coefficients.

Overall, the results of this study do not support the theoretical unidimensionality of the Kazakhstani version of the scale (Reynolds, 1982). In contrast, the findings clearly suggest that a multidimensional factorial structure and existence of a spurious factor provide better representations of the data. On the one hand this is consistent with a growing number of studies that have challenged the use of the full and short versions of the MCSDS to measure a single factor of SDB representing "need for approval" (e.g., Paulhus, 1984; Barger, 2002; Stöber et al., 2002; Leite and Beretvas, 2005). On the other hand, the significant random component along with the substantive component supports the idea that the results of MCSDS-Form C were affected by the response style of the teachers (Maydeu-Olivares and Coffman, 2006).

The results of this study suggest that both a two and a three correlated factor models demonstrated satisfactory fit to the data in the CFAs. Their more complex alternatives (i.e., bifactor and hierarchical factor models) were underidentified or demonstrated low factor loadings for some of the items. Although the three-factor model showed a relatively better performance than the two-factor model, the later seemed to provide a more empirically adequate and theoretically sound structure for the Kazakhstani version of the MCSDS – Form C. This could be due to at least four reasons. First, the EFA with oblique rotation showed substantial item cross-loadings (>0.20) for the three-factor model. Such cross-loadings present a great challenge for classical CFA, since significant cross-loadings can affect model estimation and identification (Mai et al., 2018; Zhang et al., 2021). Second, the moderate to high correlation between the second and third factors ($r < 0.56$) in the three-factor model suggests that both factors essentially represent one construct. Furthermore, the low correlation between the two components in the two-factor and three-factor CFA models ($r < 0.20$) suggests that these two are separate but related constructs. Third, the test of measurement invariance across age and language in the three-factor model showed improper solution and non-convergence issues. This can be due to the small number of indicators (i.e., two items) for factor 3. Such results are in line with findings on estimation and convergence in CFA models. For instance, Anderson and Gerbing (1984) found that the likelihood of non-convergent and improper cases increases in models with small sample sizes and a small number of indicators per factor. Similarly, Ding et al. (1995) showed that the frequency of improper solutions depends on small samples and two indicators per factor in CFA models. For the two-factor model, we did not have non-convergence and improper solutions across all groups, although we found statistical differences between men and women teachers. Fourth, the internal consistency coefficients demonstrate slightly better reliability of the scores in the two-factor solution compared to the three-factor solution. More specifically, the alpha coefficients suggest that the items of the scale are relatively accurate when measuring two dimensions, but they do not precisely measure a third dimension (α = 0.62). However, the low omega coefficients

for all subscale scores ($\omega < 0.60$) indicate that neither the two-factor nor the three-factor models offer high confidence in measuring SDB with an acceptable level of precision.

In addition to these reasons, the two-factor model also presents itself as a better solution from a theoretical point of view. **Figure 5** presents the resulting distribution of items across the two latent factors. The Kazakhstani version of the MCSDS Form C seems to resemble two separate dimensions of social desirability: attribution and denial (Millham, 1974). The former accounts for assigning socially favorable traits to oneself, while the latter represents a tendency to deny socially unfavorable traits. Furthermore, existing studies of the original MCSDB scale over the years in different cultural contexts confirmed that attribution and denial are the two underlying dimensions of the full as well as the short forms (Ramanaiah et al., 1977; Loo and Thorpe, 2000; Tao et al., 2009; He et al., 2015; Kurz et al., 2016). In this context, it can be argued that the first factor accounts for the dimension of attribution, whereas the second factor represents the dimension of denial. Individuals with high scores on both constructs, rather than being concerned with the actual meaning of their behavior, are more concerned with the external disapproving judgment (Millham, 1974). Furthermore, based on the low factor correlation ($r < 0.20$) we support the idea that these two sub concepts should be measured separately (Fischer and Fick, 1993).

Alternatively, the RIIFA model demonstrated the existence of a spurious factor associated with the item keying. In this model, the random component accounted for the substantial percentage of variance (21%), whereas the substantial factor accounted for 18%. The bigger proportion of variance of the random intercept suggests that the scale answers depend more on the method factor rather than the substantial factor that represents SDB. Thus, unlike the two-factor solution, the second dimension is not substantive and merely depicts idiosyncratic use of the scale by the teachers. Moreover, in comparison with the two-factor solution, the RIIFA model produced a relatively better fit. Overall, in this particular sample of Kazakhstani teachers, these findings present an alternative interpretation of the MCSDS-Form C results that do not support the existence of the attribution and denial dimensions. Moreover, the RIIFA results indicate low factor loadings between the substantial factor and items 6, 8, 11, 12 ($\beta = 0.16$, $\beta = 0.29$) suggesting weak relation between the items and the substantive factor, as well as the clear grouping of negatively and positively worded items.

Collectively, based on the results above, we favor the RIIFA solution and suggest interpreting the results of MCSDS-Form C as dependent on teacher response styles, not on the substantive factors representing social desirability. Still, the two-factor solution can be considered as a good hypothetical alternative that should be considered when working with MCSDS-Form C.

This is especially relevant considering some striking results in the latest TALIS 2018 study. For instance, in Kazakhstan 72% of teachers self-assessed their level of preparedness in classroom management as good and very good. In comparison, the OECD average in this component was 53% (OECD, 2019). In fact, in all items on preparedness Kazakhstani teachers indicated higher percentages of good and very good levels than their colleagues from OECD, the range of percentage difference is from 9 to 22% (Information-Analytic Center [IAC], 2019).

A plausible explanation for the high percentages of SDS in the present study is the higher number of females in the sample. In fact, the population distribution indicates a proportion of 4 to 1 (80 to 20%) in favor of female teachers (Information-Analytic Center [IAC], 2020). Previous research has shown that females tend to exhibit higher SDS than male respondents (e.g., Barger, 2002; Booth-Kewley et al., 2007; Fastame and Penna, 2012; Bossuyt and Van Kenhove, 2018). Apart from this, some broader cultural differences, such as collectivism and individualism, may lead to differences in responses. High levels of SDB in collectivist societies (e.g., like Kazakhstan) have been widely discussed in the literature (Middleton and Jones, 2000; van Hemert et al., 2002; Kim and Kim, 2016; Ryan et al., 2021). For example, van Hemert et al. (2002) found a negative correlation between the Lie scale and individualist culture. The Lie scale constitutes a part of EPQ (Eysenck Personality Questionnaire) and measures social conformity and behavior of faking good (Eysenck and Eysenck, 1991). Thus, one of the possible major reasons behind the poor reliability of the MCSDB – Form C in this study could relate to the general tendency to give dishonest answers according to collectivist cultural orientations in Kazakhstan. Unfortunately, we do not have enough evidence to further elaborate on this point since our primary interest was to check psychometric properties of the short form. Surprisingly, this article is one of the few attempts to study an instrument measuring SDB in a post-soviet country of Central Asia with collectivist culture, even though the social desirability was extensively studied cross-culturally elsewhere, across different fields of social science including but not limited to psychology, education, and sociology. Moreover, a large part of the previous research utilizing full and short forms of MCSDS was mainly concerned with social desirability as representing substantive dimensions but did not consider the potential effect of a response style on the scale answers. In this respect, when working with MCSDS forms we propose to account for both, substantive, and method factors by using traditional CFA and the RIIFA models. More research is needed in this direction. We believe that this article will open a path to future research on social desirability bias as a response pattern and as a personality characteristic with special focus on collectivist post-soviet countries of Central Asia.

Speaking about the limitations of the article, we can highlight several major factors that can potentially affect the results. First, according to the results, the scale is not a perfect measurement of social desirability; ideally, it would be appropriate to repeat the above procedure on the full MCSDB scale consisting of 33 items. This article focuses only on one of the existing short forms proposed by Nederhof (1985). The second limitation is related to the target population of the survey and its subgroups' specifics. Although the sample is representative, it focuses only on the subject teachers. Sampling issues are not new or specific to this particular Kazakhstani MCSD survey. Many studies have identified sampling representations as limitations (Beretvas et al., 2002; Sârbescu et al., 2012). Although some of these studies indicate an overwhelming participation of males (Sârbescu et al., 2012), other studies find issues of reliability differences on social desirability even with less differences in gender representation (Loo and Thorpe, 2000; Beretvas et al., 2002). Thus, future research on SDB in Kazakhstan and in

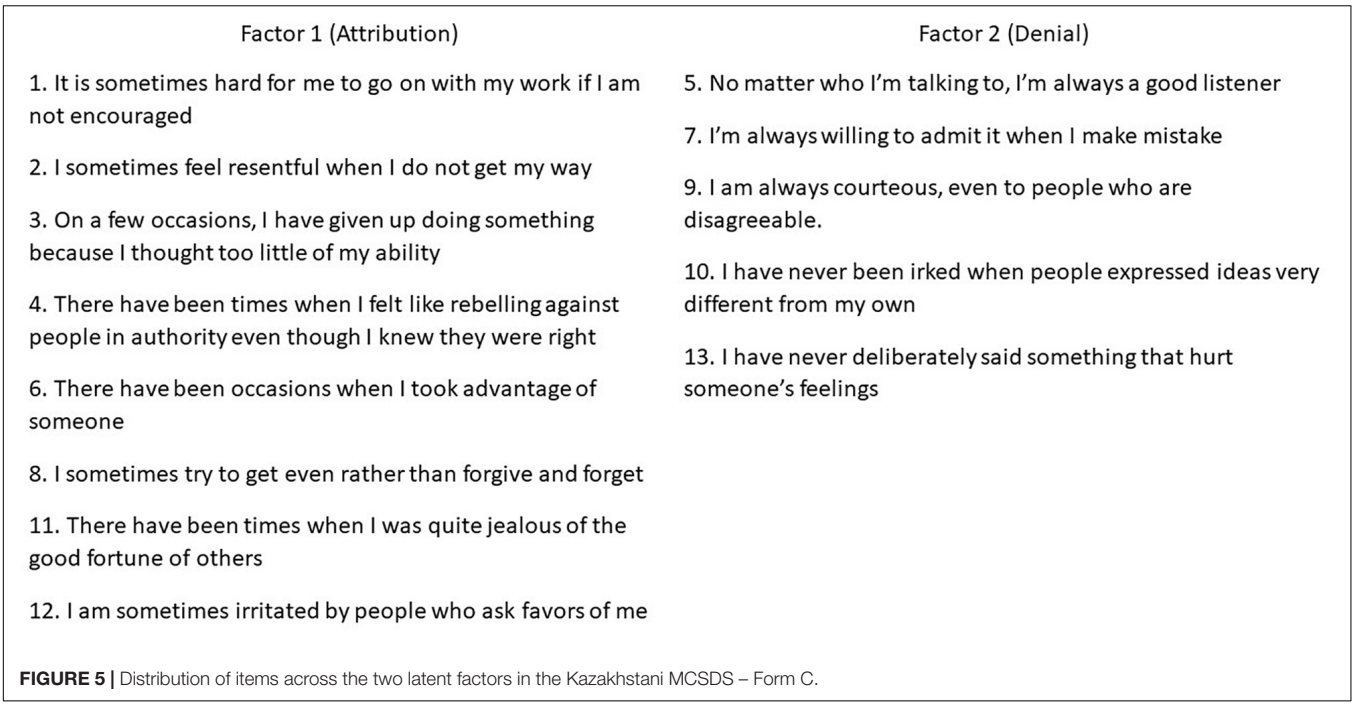

FIGURE 5 | Distribution of items across the two latent factors in the Kazakhstani MCSDS – Form C.

societies with predominantly collectivist culture can broaden the focus from specific target subpopulations to the general country-wide population testing either several short forms or the full MCSDB scale. Third, although the MCSDB scale is one of the most widely spread instruments, there are other traditional scales (Edwards, 1957; Sackheim and Gur, 1978; Paulhus, 1988; Eysenck and Eysenck, 1991) that can be used together with the MCSDB to measure social desirability and to test for convergent validity. The fourth limitation concerns measurement invariance for the RIIFA model. Although due to low factor loadings we did not calculate configural, metric and scalar invariance models nevertheless future research could include traditional MI as well as computation of a specific (factor and method) metric invariance to test whether the substantive factor and the method factor are independent (Steenkamp and Maydeu-Olivares, 2020).

In addition, factor analysis works best with the continuous data, employed in this study on the matrix of tetrachoric correlation, it is a limited information model, and the results must be regarded as an approximation of the full model (Mislevy, 1986; Schumacker and Beyerlein, 2000). Therefore, in exploring the factorial structure of MCSDS – Form C, future research can focus on full information models that allow one to work directly with categorical data and account for potentially important cross-loadings. Instead of the classical approach used in this article, one could use either Bayesian CFA or MIRT models. In the former, one can account for important cross-loadings in the model by placing normal priors with small variance on them (Muthen and Asparouhov, 2012). In the latter, MIRT models specifically work with categorical binary and polytomous items and allow estimation of within item structure where an item can be associated with several latent traits, which is not possible in classical CFA.

CONCLUSION

Research on SDB requires measurement instruments that provide reliable and valid scores in local contexts, cultures, and languages. In this study, we report several approaches to determine the psychometric performance of the Kazakhstani version of the MCSDS – Form C. We conclude that when using the Kazakhstani version of the MCSDS – Form C, if the RIIFA modes does not signal the presence of a significant method factor along with the substantive factor, then separate attribution and denial scores should be used instead of a total score measuring SDB. Furthermore, caution should be exercised when interpreting these scores due to the low omega reliability coefficients obtained for both subscales. The measurement of attribution and denial is equivalent across geographic location (urban vs. rural), language (Kazakh vs. Russian), and age groups, but these dimensions seem to be interpreted differently between male and female participants. Furthermore, MCSDS does not seem to be a perfect instrument for the context of Kazakhstani teachers because the collective culture of the Kazakhstani society combined with the current rigid vertical system of education could have an impact on the answers to the questions of the instrument. Despite these limitations, the validation of the Kazakhstani version of the MCSDS – Form C presented in this study is a first step in facilitating further research and measurement of SDB in post-Soviet Kazakhstan and other Central Asian countries.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of organizational data confidentiality

policy. Requests to access the datasets should be directed to KN.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the Local Legislation and Institutional Requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KN and DH-T contributed to the conception and design of the study, organized the database, performed the statistical analysis, and wrote the first draft of the manuscript. AA and UO wrote the sections of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

We acknowledge the support from the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan. This work was carried out within the grant OR11465485.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.822931/full#supplementary-material

REFERENCES

Anderson, J. C., and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika* 49, 155–173. doi: 10.1007/BF02294170

Ballard, D. (1992). Short forms of the Marlowe-Crowne social desirability scale. *Psychol. Rep.* 71, 1155–1160. doi: 10.2466/pr0.1992.71.3f.1155

Barendese, M. T., Oort, F. J., and Timmerman, M. E. (2014). Using exploratory factor analysis to determine dimensionality of discrete responses. *Struct. Equ. Model.* 22, 1–15. doi: 10.1080/10705511.2014.934850

Barger, S. D. (2002). The Marlowe-Crowne affair: short forms, psychometric structure, and social desirability. *J. Pers. Assess.* 97, 286–305. doi: 10.1207/S15327752JPA7902_11

Beretvas, S. N., Meyers, J. L., and Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne social desirability scale. *Educ. Psychol. Meas.* 62, 570–589. doi: 10.1177/0013164402062004003

Bernardi, R. A. (2006). Associations between Hofstede's cultural constructs and social desirability response bias. *J. Bus. Ethics* 65, 43–53. doi: 10.1007/s10551-005-5353-0

Booth-Kewley, S., Larson, G. E., and Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Comput. Hum. Behav.* 23, 463–477. doi: 10.1016/j.chb.2004.10.020

Bossuyt, S., and Van Kenhove, P. (2018). Assertiveness bias in gender ethics research: why women deserve the benefit of the doubt. *J. Bus. Ethics* 150, 727–739. doi: 10.1007/s10551-016-3026-9

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

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

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *Br. J. Math. Stat. Psychol.* 37, 62–83. doi: 10.1111/j.2044-8317.1984.tb00789.x

Carrol, J. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika* 26, 347–372. doi: 10.1007/BF02289768

Clair, J. M., and Wasserman, J. (2007). "Health and medicine," in *The Blackwell Encyclopedia of Sociology*, ed. G. Ritzer (John Wiley & Sons), 2067–2072.

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

Crowne, D. P., and Marlowe, D. (1964). *The Approval Motive: Studies in Evaluation Dependence*. New York, NY: Wiley.

Ding, L., Velicer, W. F., and Harlow, L. L. (1995). Effects of estimation methods, number of indicators per factor, and improper solutions on structural equation modeling fit indices. *Struct. Equ. Model.* 2, 119–144. doi: 10.1080/10705519509540000

DiStefano, C., and Motl, R. W. (2009). Personality correlates of method effects due to negatively worded items on the Rosenberg self-esteem scale. *Pers. Individ. Differ.* 46, 309–313. doi: 10.1016/j.paid.2008.10.020

Edwards, A. L. (1957). *The Social Desirability Variable in Personality Assessment and Research*. New York, NY: Dryden.

Eysenck, H. J., and Eysenck, S. B. G. (1991). *Manual of the Eysenck Personality Scales*. London: Hodder & Stoughton.

Falchikov, N., and Boud, D. (1989). Student self-assessment in higher education: a meta-analysis. *Rev. Educ. Res.* 59, 395–430. doi: 10.3102/00346543059004395

Fastame, M. C., and Penna, M. P. (2012). Does social desirability confound the assessment of self-reported measures of well-being and metacognitive efficiency in young and older adults? *Clin. Gerontol.* 35, 239–256. doi: 10.1080/07317115.2012.660411

Fischer, D. G., and Fick, C. (1993). Measuring social desirability: short forms of the Marlow-Crowne social desirability scale. *Educ. Psychol. Manage.* 53, 417–424. doi: 10.1177/0013164493053002011

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Adv. Methods Pract. Psychol. Sci.* 3, 484–501. doi: 10.1177/2515245920951747

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods* 9, 466–491. doi: 10.1037/1082-989X.9.4.466

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. New York, NY: John Wiley & Sons.

Green, S. B., and Yang, Y. (2015). Evaluation of dimensionality in the assessment of internal consistency reliability: coefficient alpha and omega coefficients. *Educ. Meas. Issues Pract.* 34, 14–20. doi: 10.1111/emip.12100

He, J., van de Vijver, F. J., Espinosa, A. D., Abubakar, A., Dimitrova, R., Adams, B. G., et al. (2015). Socially desirable responding: enhancement and denial in 20 countries. *Cross Cult. Res.* 49, 227–249. doi: 10.1177/1069397114552781

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

Information-Analytic Center [IAC] (2019). *Mezhdunarodnoe Issledovaniye Prepodavanya I Obuchenya TALIS-2018: Pervyie Resulaty Kazakhstana, Nacionalnyi Otchet, 1 tom [International Survey of Teaching and Learning TALIS-2018: First Results of Kazakhstan, National Report]*, Vol. 1. Nur-Sultan: Ministry of Education and Science of the Republic of Kazakhstan.

Information-Analytic Center [IAC] (2020). *Qazaqstan Respublikasi Bilim Beru Zhiuesinin Statistikasy: Ulttyq Zhinaq. [Statistics of System of Education in Kazakhstan: National Report]*. Nur-Sultan: Ministry of Education and Science of the Republic of Kazakhstan.

Information-Analytic Center [IAC] (2021). *ICT-Competency Teacher Readiness Survey. Final Report*. Nur-Sultan: Ministry of Education and Science of the Republic of Kazakhstan.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., et al. (2021). *Package 'semTools'*. Available online at: https://cran.r-project.org/web/packages/semTools/semTools.pdf (accessed June 7, 2021).

Kim, S. H., and Kim, S. (2016). National culture and social desirability bias in measuring public service motivation. *Adm. Soc.* 48, 444–476. doi: 10.1177/0095399713498749

King, M., and Bruner, G. (2000). Social desirability bias: a neglected aspect of validity testing. *Psychol. Mark.* 17, 79–103. doi: 10.1002/(SICI)1520-6793(200002)17:2<79::AID-MAR2<3.0.CO;2-0

Kurz, S. A., Drescher, C. F., Chin, E., and Johnson, L. R. (2016). Measuring social desirability across language and sex: a comparison of Marlowe-Crowne social desirability scale factor structures in English and Mandarin Chinese in Malaysia. *PsyChJournal* 5, 92–100. doi: 10.1002/pchj.124

Lalwani, A. K., Shavitt, S., and Johnson, T. (2006). What is the relation between cultural orientation and socially desirable responding? *J. Pers. Soc. Psychol.* 90, 165–178. doi: 10.1037/0022-3514.90.1.165

Le, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *J. Stat. Softw.* 25, 1–18. doi: 10.18637/jss.v025.i01

Leite, W. L., and Beretvas, S. N. (2005). Validation of scores on the Marlowe–Crowne social desirability scale and the balanced inventory of desirable responding. *Educ. Psychol. Meas.* 65, 140–154. doi: 10.1177/0013164404267285

Linting, M., Meulman, J. J., Groenen, P. J. F., and van der Kooij, A. J. (2007). Nonlinear principal component analysis: introduction and application. *Psychol. Methods* 12, 336–358. doi: 10.1037/1082-989X.12.3.336

Loo, R., and Loewen, P. (2004). Confirmatory factor analyses of scores from full and short versions of the Marlowe-Crowne social desirability scale. *J. Appl. Soc. Psychol.* 34, 2343–2352. doi: 10.1111/j.1559-1816.2004.tb01980.x

Loo, R., and Thorpe, K. (2000). Confirmatory factor analyses of the full and short versions of the Marlowe-Crowne social desirability scale. *J. Soc. Psychol.* 140, 628–635. doi: 10.1080/00224540009600503

Mai, Y., Zhang, Z., and Wen, Z. (2018). Comparing exploratory structural equation modeling and existing approaches for multiple regression with latent variables. *Struct. Equ. Model.* 25, 737–749. doi: 10.1080/10705511.2018.1444993

Mair, P., De Leeuw, J., and Groenen, J. F. P. (2019). *Package 'Gifi'*. Available online at: https://cran.r-project.org/web/packages/Gifi/Gifi.pdf (accessed June 3, 2021).

Marsh, H. (1996). Positive and negative global self-esteem: a substantively meaningful distinction or artifactors. *J. Pers. Soc. Psychol.* 70, 810–819. doi: 10.1037/0022-3514.70.4.810

Marsh, H. W. (1989). Confirmatory factor analysis of Multitrait–Multimethod data: many problems and a few solutions. *Appl. Psychol. Meas.* 13, 335–361. doi: 10.1177/014662168901300402

Maydeu-Olivares, A., and Coffman, D. L. (2006). Random intercept item factor analysis. *Psychol. Methods* 11, 344–362. doi: 10.1037/1082-989X.11.4.344

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Middleton, K. L., and Jones, J. L. (2000). Socially desirable response sets: the impact of country culture. *Psychol. Mark.* 17, 149–163. doi: 10.1002/(SICI)1520-6793(200002)17:2<149::AID-MAR6<3.0.CO;2-L

Millham, J. (1974). Two components of need for approval and their relationship to cheating following success and failure. *J. Res. Pers.* 8, 378–392. doi: 10.1016/0092-6566(74)90028-2

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *J. Educ. Stat.* 11, 3–31. doi: 10.3102/10769986011001003

Muthen, B., and Asparouhov, T. (2012). Bayesian structural equation modeling. A more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthen, O. B., Du Toit, H. C. S., and Spisic, D. (1997). *Robust Inference Using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling with Categorical and Continuous Outcomes*. Available online at: https://www.statmodel.com/download/Article_075.pdf (accessed June 15, 2021).

Nederhof, A. J. (1985). Methods of coping with social desirability bias: a review. *Eur. J. Soc. Psychol.* 15, 263–280. doi: 10.1002/ejsp.2420150303

Nieto, M. D., Garrido, L. E., Martinez-Molina, A., and Abad, F. J. (2021). Modeling wording effects does not help in recovering uncontaminated person scores: a systematic evaluation with random intercept item factor analysis. *Front. Psychol.* 12:685326. doi: 10.3389/fpsyg.2021.685326

OECD (2019). *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners, TALIS*. Paris: OECD Publishing. doi: 10.1787/1d0bc92a-en

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *J. Pers. Soc. Psychol.* 46, 598–609. doi: 10.1037/0022-3514.46.3.598

Paulhus, D. L. (1988). *Assessing Self-Deception and Impression Management in self-Reports: The Balanced Inventory of Desirable Responding*. Unpublished manual. Vancouver, BC: University of British Columbia.

Paulhus, D. L. (1991). "Measurement and control of response bias," in *Measures of Personality and Social Psychological Attitudes*, eds J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (San Diego, CA: Academic Press), 17–59. doi: 10.1016/B978-0-12-590241-0.50006-X

Paulhus, D. L., and Vazire, S. (2007). "The self-report method," in *Handbook of Research Methods in Personality Psychology*, eds R. Q. Robins, R. C. Fraley, and R. F. Krueger (New York, NY: The Guilford Press), 224–239.

Pearson, K. (1900). Mathematical contributions to the theory of evolution. – VII. On the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. Lond. Ser. A* 195, 1–47. doi: 10.1098/rsta.1900.0022

Phillips, L. D., and Clancy, J. K. (1972). Some effects of "Social Desirability" in survey studies. *Am. J. Sociol.* 77, 921–940. doi: 10.1086/225231

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available online at: https://www.R-project.org/ (accessed June 1, 2021).

Ramanaiah, N. V., Schill, T., and Leung, L. S. (1977). A test of the hypothesis about the two-dimensional nature of the Marlowe-Crowne social desirability scale. *J. Res. Pers.* 11, 251–259. doi: 10.1016/0092-6566(77)90022-8

Revelle, J. O. (2021). *Package 'psych'*. Available online at: https://cran.rstudio.org/web/packages/psych/psych.pdf (accessed June 2, 2021).

Reynolds, W. M. (1982). Development of reliable and valid short forms of the Marlowe-Crowne social desirability scale. *J. Clin. Psychol.* 38, 119–125. doi: 10.1002/1097-4679(198201)38:1<119::AID-JCLP2270380118<3.0.CO;2-I

Robins, R. W., Tracy, J. L., and Sherman, J. W. (2007). "What kinds of methods do personality psychologists use? A survey of journal editors and editorial board members," in *Handbook of Research Methods in Personality Psychology*, eds R. Q. Robins, R. C. Fraley, and R. F. Krueger (London: Guilford), 673–678.

Rosseel, Y. (2012). lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02

Ryan, A. M., Brdburn, J., Bhatia, S., Beals, E., Boyce, A. S., Martin, N., et al. (2021). In the eye of the beholder: considering culture in assessing the social desirability of personality. *J. Appl. Psychol.* 106, 452–466. doi: 10.1037/apl0000514

Sackheim, H. A., and Gur, R. C. (1978). "Self-deception, self-confrontation and consciousness," in *Consciousness and Self-Regulation: Advances in Research*, Vol. 2, eds G. E. Schwartz and D. Shapiro (New York, NY: Plenum), 139–197. doi: 10.1007/978-1-4684-2571-0_4

Sârbescu, P., Costea, I., and Rusu, S. (2012). Psychometric properties of the Marlowe-Crowne social desirability scale in a Romanian sample. *Procedia Soc. Behav. Sci.* 33, 707–711. doi: 10.1016/j.sbspro.2012.01.213

Satorra, A., and Bentler, P. M. (2000). A scaled difference Chi-square test statistic for moment structure analysis. *Psychometrika* 66, 507–514. doi: 10.1007/BF02296192

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *J. Educ. Res.* 99, 323–338. doi: 10.3200/JOER.99.6.323-338

Schumacker, R. E., and Beyerlein, S. T. (2000). Confirmatory factor analysis with different correlation types and estimation methods. *Struct. Equ. Model.* 7, 629–636. doi: 10.1207/S15328007SEM0704_6

Seol, H. (2007). A psychometric investigation if the Marlowe-Crowne social desirability scale using Rasch measurement. *Meas. Eval. Couns. Dev.* 40, 155–168. doi: 10.1080/07481756.2007.11909812

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

Steenkamp, J. B. E., and Maydeu-Olivares, A. (2020). An Updated paradigm for evaluating measurement invariance incorporating common method variance and its assessment. *J. Acad. Mark. Sci.* 49, 5–29. doi: 10.1007/s11747-020-00745-z

Stöber, J., Dette, D. E., and Musch, J. (2002). Comparing continuous and dichotomous scoring of the balanced inventory of desirable responding. *J. Pers. Assess.* 78, 370–389. doi: 10.1207/S15327752JPA7802_10

Strahan, G., and Gerbasi, C. K. (1972). Short, homogenous versions of the Marlowe-Crowne social desirability scale. *J. Clin. Psychol.* 28, 191–193. doi: 10.1002/1097-4679(197204)28:2<191::AID-JCLP2270280220<3.0.CO;2-G

Tao, P., Guoying, D., and Brody, S. (2009). Preliminary study of a Chinese language short form of the Marlowe-Crowne social desirability scale. *Psychol. Rep.* 105, 1039–1046. doi: 10.2466/PR0.105.F.1039-1046

The Agency on Statistics of the Republic of Kazakhstan (2011). *Results of the 2009 National Population Census of the Republic of Kazakhstan: Analytical Report*. Available online at: https://stat.gov.kz/ (accessed July 19, 2021).

Tracey, T. J. G. (2016). A note on socially desirable responding. *J. Couns. Psychol.* 63, 224–232. doi: 10.1037/cou0000135

UNESCO (2011). *UNESCO ICT Competency Framework for Teachers*. Paris: UNESCO.

van de Mortel, T. F. (2008). Faking it: social desirability response bias in self-report research. *Aust. J. Adv. Nurs.* 25, 40–48.

van Hemert, A. D., van de Vijver, F. J. R., Poortinga, H. Y., and Georgas, J. (2002). Structural and functional equivalence of the Eysenck personality questionnaire within and between countries. *Pers. Individ. Differ.* 33, 1229–1249. doi: 10.1016/S0191-8869(02)00007-7

Ventimiglia, M., and MacDonald, D. A. (2012). An examination of the factorial dimensionality of the Marlowe Crowne social desirability scale. *Pers. Individ. Differ.* 52, 487–491. doi: 10.1016/j.paid.2011.11.016

Verardi, S., Dahourou, D., Ah-Kion, J., Bhowon, U., Tseung, C. N., Amoussou-Yeye, D., et al. (2009). Psychometric properties of the Marlowe-Crowne social desirability scale in eight African countries and Switzerland. *J. Cross Cult. Psychol.* 41, 19–34. doi: 10.1177/0022022109348918

Vésteinsdóttir, V., Reips, U. D., Joinson, A., and Thorsdottir, F. (2015). Psychometric properties of measurements obtained with the Marlowe-Crowne social desirability scale in an Icelandic probability based internet sample. *Comput. Hum. Behav.* 49, 608–614. doi: 10.1016/j.chb.2015.03.044

Vésteinsdóttir, V., Reips, U. D., Joinson, A., and Thorsdottir, F. (2017). An item level evaluation of the Marlowe-Crowne social desirability scale using item response theory on Icelandic internet panel data and cognitive interviews. *Pers. Individ. Differ.* 107, 164–173. doi: 10.1016/j.paid.2016.11.023

Winter, L., Hernández-Torrano, D., McLellan, R., Almukhambetova, A., and Brown-Hajdukova, E. (2020). A contextually adapted model of school engagement in Kazakhstan. *Curr. Psychol.* doi: 10.1007/s12144-020-00758-5 [Epub ahead of print].

Zhang, B., Luo, J., Sun, T., Cao, M., and Drasgow, F. (2021). Small but nontrivial: a comparison of six strategies to handle cross-loadings in bifactor predictive models. *Multivariate Behav. Res.* doi: 10.1080/00273171.2021.1957664 [Epub ahead of print].

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 | Frontiers in Psychology

METHODS
published: 25 April 2022
doi: 10.3389/fpsyg.2022.841372

Bayesian Analysis of Aberrant Response and Response Time Data

Zhaoyuan Zhang[1,2], *Jiwei Zhang*[3*] and *Jing Lu*[4*]

[1] School of Mathematics and Statistics, Yili Normal University, Yining, China, [2] Institute of Applied Mathematics, Yili Normal University, Yining, China, [3] School of Mathematics and Statistics, Yunnan University, Kunming, China, [4] Key Laboratory of Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China

In this article, a highly effective Bayesian sampling algorithm based on auxiliary variables is proposed to analyze aberrant response and response time data. The new algorithm not only avoids the calculation of multidimensional integrals by the marginal maximum likelihood method but also overcomes the dependence of the traditional Metropolis–Hastings algorithm on the tuning parameter in terms of acceptance probability. A simulation study shows that the new algorithm is accurate for parameter estimation under simulation conditions with different numbers of examinees, items, and speededness levels. Based on the sampling results, the powers of the two proposed Bayesian assessment criteria are tested in the simulation study. Finally, a detailed analysis of a high-state and large-scale computerized adaptive test dataset is carried out to illustrate the proposed methodology.

Keywords: aberrant responses, Bayesian inference, mixture hierarchical model, Pólya-gamma distribution, rapid guessing behavior, Gibbs sampling algorithm

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Pablo Gomez,
California State University, United States
Dylan Molenaar,
University of Amsterdam, Netherlands
Kaiwen Man,
University of Alabama, United States

***Correspondence:**
Jiwei Zhang
zhangjw713@nenu.edu.cn
Jing Lu
luj282@nenu.edu.cn

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

Received: *22 December 2021*
Accepted: *15 March 2022*
Published: *25 April 2022*

Citation:
Zhang Z, Zhang J and Lu J (2022) Bayesian Analysis of Aberrant Response and Response Time Data. Front. Psychol. 13:841372. doi: 10.3389/fpsyg.2022.841372

1. INTRODUCTION

In educational psychological assessments, examinees often perform different types of test-taking behaviors (Bolt et al., 2002; Boughton and Yamamoto, 2007; Goegebeur et al., 2008; Chang et al., 2014; Wang and Xu, 2015; Wang et al., 2018; Man et al., 2018; Man and Harring, 2021). One is the solution behavior, in which the examinee gives a response after careful consideration to each part of an item (Schnipke and Scrams, 1997; Bolt et al., 2002; Wise and Kong, 2005; Wang and Xu, 2015). An alternative is the rapid guessing behavior, in which the examinee simply seeks to obtain an answer quickly without a deep thought process; this behavior often occurs in high-stakes tests owing to insufficient time and in low-stakes tests owing to lack of motivation. In fact, the traditional item response theory (IRT) model is based on the assumption that the correct response probability increases with the ability of the test taker under the solution behavior. The correct response probability under the rapid guessing behavior is actually rarely dependent on the measure constructed by the test (Lord and Novick, 1968; Wise and DeMars, 2006; Boughton and Yamamoto, 2007; Goegebeur et al., 2008). Numerous studies have shown that the presence of rapid guessing behavior inevitably leads to biased inferences of the item and person parameters (Bolt et al., 2002; Wise and DeMars, 2006; Boughton and Yamamoto, 2007; Goegebeur et al., 2008; Chang et al., 2014; Wang and Xu, 2015; Wang et al., 2018). Therefore, appropriate models need to be constructed to capture both solution behavior and rapid guessing behavior to reduce these biased parameter estimates. Before we analyze aberrant response behavior, we provide an explanation of the change point, which is the cut-off point at which an examinee adopts different response strategies. By considering a change point, Bolt et al. (2002) classified examinees in the speeded group before the

change point and found that they were more likely to adopt the solution behavior, whereas examinees who transferred from the speeded group to the non-speeded group after the change point were more likely to choose the rapid guessing behavior. In contrast to models using fixed change point locations, Boughton and Yamamoto (2007) proposed the more flexible HYBRID model, which allowed different examinees to have change points at different locations. The model assumes that examinees' responses follow a Rasch model until a particular point in a given examinee's test, after which the responses to all items are randomly guessed. Goegebeur et al. (2008) proposed a speeded model with one change point to characterize the gradual switch between response strategies by introducing an additional examinee-specific change-rate parameter. In addition, Wise and DeMars (2006) proposed an effort-moderated IRT model to decompose the correct response probability into a mixture of two sub-models. The two sub-models were used to characterize the solution behavior and rapid guessing behavior, respectively.

In parallel with the abovementioned item response data, response time, which is an important type of important auxiliary information, has been widely used to distinguish between two different behaviors (Schnipke and Scrams, 1997; Wise and DeMars, 2006; van der Linden and Guo, 2008; Wang and Xu, 2015). van der Linden and Guo (2008) found that examinees' response times in a high-stakes achievement test showed a mixture of two different distributions. Similarly, Schnipke and Scrams (1997) verified that the distribution of response times for end-of-test items showed a bimodal pattern in a high-stakes exam. In the study of (Schnipke and Scrams, 1997), a two-state mixture model was proposed to decompose the distribution of response times for each item into two parts. The two parts of the response times quantified the solution behavior and the rapid guessing behavior, respectively. Wang and Xu (2015) proposed a mixture model to consider differences between item responses and response time patterns resulting from the solution behavior and rapid guessing behavior. The mixture model used both item response and response time information and considered multiple switch points for each examinee.

A variety of estimation methods have been proposed to estimate the parameters of the IRT and response time models. In the frequentist framework, the most common method is the marginal maximum likelihood estimation (MMLE) *via* expectation maximization algorithm (Bock and Aitkin, 1981; Baker and Kim, 2004). However, the main drawback of marginal maximum likelihood methods is the inevitable need for tedious approximation of the multidimensional integral using numerical integration (Bock and Schilling, 1997; Rabe-Hesketh et al., 2002, 2005) or Monte Carlo integration (Kuk, 1999; Skaug, 2002) when the latent variables are high dimensional. This is because the number of discrete quadrature points required increases exponentially as the number of latent variables increases linearly during the computation (Converse et al., 2021, p. 1465). Although the adaptive quadrature method has been used to deal with the computational deficiency by using a small number of quadrature points, the problem has not been completely solved (Jiang and Templin, 2019). In addition, the comparison method of the MMLE is simplistic; comparison methods other than the root mean square error of approximation are seldom used (Zhang et al., 2019). Compared with the MMLE method, first, the Bayesian method allows one to update knowledge by using proper informative priors based on previous studies, the posterior distribution being more precise than the likelihood or the prior alone (Jackman, 2009). The incorporation of proper informative priors into the Bayesian analysis can be used to obtain better results in the case of small to moderate sample sizes. In addition, even if weakly informative inaccurate priors are used, the performance of the Bayesian method does not deteriorate. Second, Bayesian estimation does not rely on asymptotic arguments and can give more reliable results for small samples (Lee and Song, 2004; Song and Lee, 2012). Third, another major advantage of Bayesian analysis is the ability to analyze models that are computationally heavy or impossible to estimate with MMLE. These include models with categorical outcomes with many latent variables and, thus, many dimensions of numerical integrations (Asparouhov and Muthén, 2010b; Muthén, 2010).

In the current study, an efficient Pólya–gamma Gibbs sampling algorithm (Polson et al., 2013) in a fully Bayesian framework is proposed to estimate the parameters of the mixture model of Wang and Xu (2015). Compared with traditional Bayesian sampling algorithms, e.g., the Metropolis–Hastings sampling algorithm (Metropolis et al., 1953; Hastings, 1970; Tierney, 1994; Chib and Greenberg, 1995; Chen et al., 2000), Gibbs sampling algorithm (Geman and Geman, 1984; Tanner and Wong, 1987; Gelfand and Smith, 1990; Albert, 1992; Béguin and Glas, 2001; Fox and Glas, 2001), and the advantages of the Pólya–gamma Gibbs sampling algorithm are presented from multiple perspectives. First, the Pólya–gamma Gibbs sampling algorithm avoids retrospective tuning in the Metropolis–Hastings sampling algorithm if we do not know how to choose a proper tuning parameter or if no value for the tuning parameter is appropriate. It always keeps the drawn samples accepted, thereby increasing the sampling efficiency (Zhang et al., 2020). Second, the Pólya–gamma Gibbs sampling algorithm can transform the non-conjugate model into the conjugate model by using augmented auxiliary variables. With the help of the traditional Gibbs sampling algorithm, posterior sampling is easier to implement (Polson et al., 2013). Third, in Bayesian estimation, prior distributions of model parameters and observed data likelihood produce a joint posterior distribution for the model parameters. The prior specifications and prior sensitivity are important aspects of Bayesian inference (Ghosh and A. Ghosh, 2000). In fact, the Pólya–gamma Gibbs sampling algorithm is not sensitive to the specification of the prior distribution. It can still obtain satisfactory results even if the proper or mis-specification priors are adopted (Zhang et al., 2020).

The rest of this article is organized as follows. Section 2 contains an introduction to the mixture hierarchical model and the corresponding identification restrictions. A detailed implementation of the Pólya–gamma Gibbs sampling algorithm is described in Section 3. In Section 4, two simulations focus on the parameter recovery performance of the Bayesian algorithm using the results of the model assessments. In addition, the quality of the Bayesian algorithm is investigated using

high-state and large-scale computerized adaptive test data in Section 5. We conclude the article with a brief discussion in Section 6.

2. MODELS

Following Wang and Xu (2015), the mixture model is used to distinguish solution behavior from rapid guessing behavior. The correct response probability of examinee i on item j is assumed to follow a mixture decomposition

$$P\left(Y_{ij}=1\left|\eta_{ij}\right.\right) = \left(1-\eta_{ij}\right)P\left(Y_{ij}=1\left|\eta_{ij}=0\right.\right) + \eta_{ij}P\left(Y_{ij}=1\left|\eta_{ij}=1\right.\right),$$

where η_{ij} is a latent response behavior indicator variable, $\eta_{ij}=1$ denotes the case where examinee i answers item j by rapid guessing behavior, and $\eta_{ij}=0$ denotes the solution behavior. $P\left(Y_{ij}=1\left|\eta_{ij}=0\right.\right)$ quantifies the probability of a correct response resulting from the solution behavior, whereas $P\left(Y_{ij}=1\left|\eta_{ij}=1\right.\right)$ captures the probability of a correct response with the rapid guessing behavior. We use the two-parameter logistic (2PL; Birnbaum, 1968) model for the solution behavior,

$$P\left(Y_{ij}=1\left|\eta_{ij}=0,\theta_i,a_j,b_j\right.\right)=\frac{\exp\left[a_j\left(\theta_i-b_j\right)\right]}{1+\exp\left[a_j\left(\theta_i-b_j\right)\right]},$$

where a_j and b_j are the discrimination and difficulty parameters of item j, and θ_i denotes the ability of the $i$the examinee. The probability that examinee i answers item j correctly by the rapid guessing behavior is g_j; this is an item-specific probability:

$$P\left(Y_{ij}=1\left|\eta_{ij}=1\right.\right)=g_j.$$

In parallel with the mixture item response model, the observed response time T_{ij}^{obs} is

$$T_{ij}^{obs}=\left(1-\eta_{ij}\right)T_{ij}+\eta_{ij}C_{ij},$$

where T_{ij} represents the time required for examinee i to respond to item j using solution behavior, and C_{ij} represents the time required for examinee i to respond to item j using rapid guessing behavior. Therefore, given latent indicator variable η_{ij}, the density function of observed response time T_{ij}^{obs} can be denoted as

$$p_{ij}\left(t_{ij}\left|\eta_{ij}\right.\right)=\left(1-\eta_{ij}\right)f_{ij}\left(t_{ij}\right)+\eta_{ij}h_{ij}\left(t_{ij}\right),$$

where f and h represent corresponding density functions of T_{ijv} and C_{ijv}.

Response times on test items have been modeled in various families of distributions in psychometric applications, including exponential (Scheiblechner, 1979), gamma (Maris, 1993), Weibull (Rouder et al., 2003), log-normal race (Rouder et al., 2015), and semi-parametric models (Wang et al., 2013). Response time data are non-negative, and their distributions tend to be positively skewed. The log transformation would move positively skewed distributions toward symmetric shapes. We chose the log-normal distribution (van der Linden, 2006) for response times with solution behavior:

$$\log\left(T_{ij}\right)=\lambda_j-\tau_i+e_{ij},\ \ e_{ij}\sim N\left(0,\sigma_j^2\right),$$

where λ_j is the time intensity of item j; a higher value of λ_j indicates that the item is expected to consume more time. τ_i is a speed parameter of examinee i; a higher value of τ_i means that the examinee works faster and a lower response time is expected. σ_j^2 allows for differences between the variances of log-times on different items. Following the "common-guessing" (Schnipke and Scrams, 1997), the response times of the guessing behavior have a common log-normal distribution

$$\log\left(C_{ij}\right)\sim N\left(\mu_c,\sigma_c^2\right).$$

To capture across-person relationships between speed and accuracy, we assume that the ability and speed parameters have a bivariate normal distribution, to explore whether examinees with higher ability tend to answer items faster, i.e.,

$$\boldsymbol{\xi}_i=\left(\theta_i,\tau_i\right)^{'}\sim N\left(\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P\right),$$

with mean vector

$$\boldsymbol{\mu}_P=\left(\mu_\theta,\mu_\tau\right)^{'}$$

and covariance matrix

$$\boldsymbol{\Sigma}_P=\begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta\tau}\\ \sigma_{\tau\theta} & \sigma_\tau^2\end{pmatrix}.$$

2.1. Model Identification

In the 2PL model, to eliminate the trade-off between ability θ and difficulty parameter b in location, we only need to fix the mean population level of ability to zero. That is, $\mu_\theta=0$. To eliminate the trade-off between ability θ and discrimination parameter a in scale, we need to restrict the variance population level of ability to one. That is, $\sigma_\theta^2=1$. For the response time model with the solution behavior, to eliminate the trade-off between speed parameter τ and time intensity parameter λ in location, we need to fix the mean population level of speed to zero. That is, $\mu_\tau=0$.

There are several widely used identification restriction methods for two-parameter IRT models. The identification restrictions of our models are based on the following methods.

(1) Fix the mean population level of ability to zero and the variance population level of ability to one (Lord and Novick, 1968; Bock and Aitkin, 1981; Fox and Glas, 2001; Fox, 2010). That is, $\theta\sim N(0,1)$.
(2) Restrict the sum of item difficulty parameters to zero and the product of item discrimination parameters to one (Fox and Glas, 2001; Fox, 2005, 2010). That is,

$$\sum_{j=1}^{J}b_j=0\quad\text{and}\quad\prod_{j=1}^{J}a_j=1.$$

(3) Fix the item difficulty parameter to a specific value, most often zero and restrict the discrimination parameter to a specific value, most often one (Fox and Glas, 2001; Fox, 2010). That is, $b_1 = 0$ and $a_1 = 1$.

3. BAYESIAN ESTIMATION USING MCMC SAMPLING

Let $\boldsymbol{\Omega} = (\eta_{ij}, , \theta_i, a_j, b_j, \lambda_j, \tau_i, \sigma_j^2, \mu_a, \sigma_a^2, \mu_b, \sigma_b^2, \mu_\lambda, \sigma_\lambda^2, \mu_c, \sigma_c^2, g_j, \sigma_{\theta\tau}, \sigma_\tau^2, \pi_i)$; then, the full joint posterior of person and item parameters given $\boldsymbol{Y}$, $\boldsymbol{T}$, and $\boldsymbol{\eta}$ is

$$
\begin{aligned}
&L\left(\boldsymbol{\Omega}\left|\boldsymbol{Y}, \boldsymbol{T}\right.\right) \\
&=\prod_{i=1}^{N}\prod_{j=1}^{J}\left[\pi_i g_j h\left(t_{ij};\, \mu_c, \sigma_c^2\right)\right]^{\eta_{ij}\cdot Y_{ij}}\left[\pi_i\left(1-g_j\right) h\left(t_{ij};\, \mu_c, \sigma_c^2\right)\right]^{\eta_{ij}\cdot\left(1-Y_{ij}\right)} \\
&\times\left[\left(1-\pi_i\right) P\left(Y_{ij}=1\left|\eta_{ij}=0, a_j, b_j, \theta_i\right.\right) f\left(t_{ij};\, \lambda_j, \tau_i, \sigma_j^2\right)\right]^{\left(1-\eta_{ij}\right).Y_{ij}} \\
&\times\left[\left(1-\pi_i\right) P\left(Y_{ij}=0\left|\eta_{ij}=0, a_j, b_j, \theta_i\right.\right) f\left(t_{ij};\, \lambda_j, \tau_i, \sigma_j^2\right)\right]^{\left(1-\eta_{ij}\right).\left(1-Y_{ij}\right)} \\
&\times p\left(\theta_i, \tau_i;\, \boldsymbol{\mu}_p, \Sigma_p\right) p\left(a_j\right) p\left(b_j\right) p\left(\lambda_j\right) p\left(\boldsymbol{\mu}_p, \Sigma_p\right), \qquad (1)
\end{aligned}
$$

where π_i is the probability that examinee i uses the rapid guessing behavior, i.e., $\pi_i = P\left(\eta_{ij}=1\right)$.

3.1. Pólya–Gamma Gibbs Sampling Algorithm

Polson et al. (2013) proposed a new data augmentation strategy for fully Bayesian inference in logistic regression. This data augmentation approach used a new class of Pólya–gamma distribution, in contrast to the data augmentation algorithm of Albert and Chib (1993), which was based on a truncated normal distribution. Here, we introduce the Pólya–gamma distribution.

Definition: Let $\{B_k\}_{k=1}^{+\infty}$ be an *independent and identically distributed* random variable sequence from a gamma distribution with parameters β and 1. That is, $B_k \sim$ gamma $(\beta, 1)$. A random variable W follows a Pólya–gamma distribution with parameters $\beta > 0$ and $\varrho \in R$, denoted $W \sim \mathrm{PG}(\beta, \varrho)$, if

$$
W \overset{D}{=} \frac{1}{2\pi}\sum_{k=1}^{+\infty}\frac{B_k}{\left(k-\frac{1}{2}\right)^2+\frac{\varrho^2}{4\pi^2}},
$$

where $\overset{D}{=}$ denotes equality in distribution. In fact, the Pólya–gamma distribution is an infinite mixture of gamma distributions, which provides the ability to sample from gamma distributions.

Based on Theorem 1 of Polson et al. (2013, page 1341, Equation 7), the likelihood contribution of the ith examinee answering the jth item under the solution behavior category $\eta_{ij} = 0$ can be expressed as

$$
\begin{aligned}
L\left(a_j, b_j, \theta_i\right) &= \frac{\left\{\exp\left[a_j\left(\theta_i-b_j\right)\right]\right\}^{Y_{ij}}}{1+\left\{\exp\left[a_j\left(\theta_i-b_j\right)\right]\right\}} \propto \exp\left\{k_{ij}\left[a_j\left(\theta_i-b_j\right)\right]\right\} \\
&\times\int_0^{\infty}\exp\left\{-\frac{W_{ij}\left[a_j\left(\theta_i-b_j\right)\right]^2}{2}\right\} p\left(W_{ij}\left|1,0\right.\right) dW_{ij},
\end{aligned} \qquad (2)
$$

where $k_{ij} = Y_{ij} - \frac{1}{2}$. $p\left(W_{ij}\left|1,0\right.\right)$ is the conditional density of W_{ij}. That is, $W_{ij} \sim \mathrm{PG}(1,0)$. The auxiliary variable W_{ij} follows a Pólya–gamma distribution with parameters (1, 0). Within the solution behavior category $\eta_{ij} = 0$, the full conditional distribution of $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}$ given the auxiliary variables, $\boldsymbol{W}$ can be written as

$$
\begin{aligned}
&p\left(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\theta}\left|\boldsymbol{\eta}, \boldsymbol{W}, \boldsymbol{Y}\right.\right) \\
&\propto\left\{\prod_{i=1}^{N}\prod_{j=1}^{J}\left[\exp\left\{k_{ij}\left[a_j\left(\theta_i-b_j\right)\right]\right\}\exp\left[-\frac{W_{ij}\left[a_j\left(\theta_i-b_j\right)\right]^2}{2}\right]\right]\right\}^{\mathrm{I}(\eta_{ij}=0)} \\
&\times\left\{\prod_{i=1}^{N} p\left(\theta_i\left|\tau_i, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P\right.\right)\right\}^{\mathrm{I}(\eta_{ij}=0)}\left\{\prod_{j=1}^{J}\left[p\left(a_j\right) p\left(b_j\right)\right]\right\}^{\mathrm{I}(\eta_{ij}=0)},
\end{aligned} \qquad (3)
$$

where $p\left(a_j\right)$ and $p\left(b_j\right)$ are the prior distributions for a_j and b_j. It is known that there are relationships between the latent ability and speed parameter, which can be constructed by a bivariate normal prior distribution $\begin{pmatrix}\theta_i\\ \tau_i\end{pmatrix} \sim N\left(\begin{pmatrix}\mu_\theta\\ \mu_\tau\end{pmatrix}, \boldsymbol{\Sigma}_P\right)$. Therefore, the conditional prior distribution of θ_i is the normal distribution

$$
\theta_i\left|\tau_i, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P\right. \sim N\left(\mu_{\theta|\tau}, \sigma_{\theta|\tau}^2\right),
$$

where $\mu_{\theta|\tau} = \mu_\theta + \sigma_{\theta\tau}\sigma_\tau^{-2}\left(\tau_i-\mu_\tau\right)$ and $\sigma_{\theta|\tau}^2 = \sigma_\theta^2 - \sigma_{\theta\tau}\sigma_\tau^{-2}\sigma_{\tau\theta}$.

Step 1: Sample the auxiliary variable W_{ij}, within the solution behavior category $\eta_{ij} = 0$, given the item discrimination and difficulty parameters a_j, b_j and the ability θ_i. According to Equation (1), the full conditional posterior distribution of the random auxiliary variable W_{ij} is given by

$$
f\left(W_{ij}\left|a_j, b_j, \theta_i\right.\right) \propto \exp\left[-\frac{W_{ij}\left[a_j\left(\theta_i-b_j\right)\right]^2}{2}\right] p\left(W_{ij}\left|1,0\right.\right).
$$

According to Biane et al. (2001) and Polson et al. (2013; p. 1341), the density function $p\left(W_{ij}\left|1,0\right.\right)$ can be written as

$$
p\left(W_{ij}\left|1,0\right.\right) = \sum_{\nu=0}^{\infty}(-1)^{\nu}\frac{(2k+1)}{\sqrt{2\pi W_{ij}}}\exp\left[-\frac{(2k+1)^2}{8W_{ij}}\right].
$$

Therefore, $f\left(W_{ij}\left|a_j, b_j, \theta_i\right.\right)$ is proportional to

$$
\sum_{\nu=0}^{\infty}(-1)^{\nu}\frac{(2k+1)}{\sqrt{2\pi W_{ij}}}\exp\left[-\frac{(2k+1)^2}{8W_{ij}}-\frac{W_{ij}\left[a_j\left(\theta_i-b_j\right)\right]^2}{2}\right].
$$

Finally, the specific form of the full conditional distribution of W_{ij} is as follows:

$$W_{ij} \sim \mathrm{PG}\left(1, \left|a_j\left(\theta_i - b_j\right)\right|\right).$$

Next, Gibbs samplers are used to draw the item parameters.

Step 2: Sample the discrimination parameter a_j for each item j. The prior distribution of a_j is assumed to follow a truncated normal distribution, i.e., $a_j \sim N\left(\mu_a, \sigma_a^2\right) \mathrm{I}\left(a_j > 0\right)$. Given $\boldsymbol{Y}$, $\boldsymbol{W}$, b_j, and $\boldsymbol{\theta}$, the fully conditional posterior distribution of a_j is given by

$$p\left(a_j \left| \boldsymbol{Y}, \boldsymbol{W}, b_j, \boldsymbol{\theta}\right.\right) \propto \prod_{i=1}^{N} \left\{ \frac{\left\{\exp\left[a_j\left(\theta_i - b_j\right)\right]\right\}^{Y_{ij}}}{1 + \exp\left[a_j\left(\theta_i - b_j\right)\right]} f\left(W_{ij} \left| a_j, b_j, \theta_i\right.\right) \right\} p\left(a_j\right),$$

where $f\left(W_{ij} \left| a_j, b_j, \theta_i\right.\right)$ is given by the following equation (for details, refer to Polson et al., 2013; p. 1341):

$$f\left(W_{ij} \left| a_j, b_j, \theta_i\right.\right) = \left\{\cosh\left(2^{-1}\left|a_j\left(\theta_i - b_j\right)\right|\right)\right\} \frac{2^0}{\Gamma(1)} \times \sum_{v=0}^{\infty} (-1)^v \frac{(2k+1)}{\sqrt{2\pi W_{ij}}} \times \exp\left[-\frac{(2k+1)^2}{8W_{ij}} - \frac{W_{ij}\left[a_j\left(\theta_i - b_j\right)\right]^2}{2}\right].$$

After rearrangement, the full conditional posterior distribution of a_j can be written as follows:

$$p\left(a_j \left| \boldsymbol{Y}, \boldsymbol{W}, b_j, \boldsymbol{\theta}\right.\right) \propto \prod_{i=1}^{N} \left\{ \frac{\left\{\exp\left[a_j\left(\theta_i - b_j\right)\right]\right\}^{Y_{ij}}}{1 + \exp\left[a_j\left(\theta_i - b_j\right)\right]} \times \left[\cosh\left(2^{-1}\left|a_j\left(\theta_i - b_j\right)\right|\right)\right] \times \exp\left[-\frac{\left[a_j\left(\theta_i - b_j\right)\right]^2 W_{ij}}{2}\right] \right\} p\left(a_j\right).$$

Therefore, the fully conditional posterior distribution of a_j follows a normal distribution truncated at 0 with mean

$$\mathrm{Var}_{a_j} \times \left(\mu_a \sigma_a^{-2} + \left[\sum_{i=1}^{N} W_{ij}\left(\theta_i - b_j\right)^2\right] \left\{ \frac{\left[\sum_{i=1}^{N}\left(1 - 2Y_{ij}\right)\left(\theta_i - b_j\right)\right]}{2\left[\sum_{i=1}^{N} W_{ij}\left(\theta_i - b_j\right)^2\right]} \right\} \right)$$

and variance

$$\mathrm{Var}_{a_j} = \left\{ \sigma_a^{-2} + \left[\sum_{i=1}^{N} W_{ij}\left(\theta_i - b_j\right)^2\right] \right\}^{-1}.$$

Step 3: Sample the difficulty parameter b_j for each item j. The prior distribution of b_j is assumed to follow a normal distribution with mean μ_b and σ_b^2. That is, $b_j \sim N\left(\mu_b, \sigma_b^2\right)$. Similarly, given $\boldsymbol{Y}$, $\boldsymbol{W}$, a_j, and $\boldsymbol{\theta}$, the fully conditional posterior distribution of b_j is given by

$$p\left(b_j \left| \boldsymbol{Y}, \boldsymbol{W}, a_j, \boldsymbol{\theta}\right.\right) \propto \prod_{i=1}^{N} \left\{ \frac{\left\{\exp\left[a_j\left(\theta_i - b_j\right)\right]\right\}^{Y_{ij}}}{1 + \exp\left[a_j\left(\theta_i - b_j\right)\right]} \times \left[\cosh\left(2^{-1}\left|a_j\left(\theta_i - b_j\right)\right|\right)\right] \times \exp\left[-\frac{\left[a_j\left(\theta_i - b_j\right)\right]^2 W_{ij}}{2}\right] \right\} p\left(b_j \left| \mu_b, \sigma_b^2\right.\right).$$

Therefore, the fully conditional posterior distribution of b_j follows a normal distribution with mean

$$\mathrm{Var}_{b_j} \times \left(\mu_b \sigma_b^{-2} + \sum_{i=1}^{N}\left[a_j^2 W_{ij}\right] \left(\frac{\sum_{i=1}^{N}\left(2a_j^2\theta_i W_{ij} - 2Y_{ij}a_j + a_j\right)}{2\sum_{i=1}^{N}\left[a_j^2 W_{ij}\right]} \right) \right)$$

and variance

$$\mathrm{Var}_{b_j} = \left\{ \sigma_b^{-2} + \sum_{i=1}^{N}\left[a_j^2 W_{ij}\right] \right\}^{-1}$$

Step 4: Sample the ability parameter θ_i for each examinee i. The conditional prior distribution of θ_i is assumed to follow a normal distribution with mean $\mu_{\theta|\tau} = \mu_\theta + \sigma_{\theta\tau}\sigma_\tau^{-2}\left(\tau_i - \mu_\tau\right)$ and $\sigma^2_{\theta|\tau} = \sigma^2_\theta - \sigma_{\theta\tau}\sigma_\tau^{-2}\sigma_{\tau\theta}$. That is, $\theta_i \sim N\left(\mu_{\theta|\tau}, \sigma^2_{\theta|\tau}\right)$. Given $\boldsymbol{Y}$, $\boldsymbol{W}$, $\boldsymbol{a}$ and $\boldsymbol{b}$, the fully conditional posterior distribution of θ_i is given by

$$p\left(\theta_i \left| \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}\right.\right) \propto \prod_{j=1}^{J} \left\{ \frac{\left\{\exp\left[a_j\left(\theta_i - b_j\right)\right]\right\}^{Y_{ij}}}{1 + \exp\left[a_j\left(\theta_i - b_j\right)\right]} \left[\cosh\left(2^{-1}\left|a_j\left(\theta_i - b_j\right)\right|\right)\right] \times \exp\left[-\frac{\left[a_j\left(\theta_i - b_j\right)\right]^2 W_{ij}}{2}\right] \right\} p\left(\theta_i \left| \mu_{\theta|\tau}, \sigma^2_{\theta|\tau}\right.\right).$$

Therefore, the fully conditional posterior distribution of θ_i follows a normal distribution with mean

$$\mathrm{Var}_{\theta_i} \times \left(\mu_{\theta|\tau}\,\sigma^{-2}_{\theta|\tau} + \sum_{j=1}^{J}\left[a_j^2 W_{ij}\right] \left(\frac{\sum_{j=1}^{J}\left(2Y_{ij}a_j + 2a_j^2 b_j W_{ij} - a_j\right)}{2\sum_{j=1}^{J}\left[a_j^2 W_{ij}\right]} \right) \right)$$

and variance

$$\mathrm{Var}_{\theta_i} = \left\{ \sigma^{-2}_{\theta|\tau} + \sum_{j=1}^{J}\left[a_j^2 W_{ij}\right] \right\}^{-1}.$$

Step 5: Sample the response behavior variable η_{ij}. The fully conditional posterior distribution of η_{ij} is a Bernoulli distribution with success probability

$$\frac{\pi_i g_j h\left(T_{ij};\, \mu_c, \sigma_c^2\right)}{\pi_i g_j h\left(T_{ij};\, \mu_c, \sigma_c^2\right) + (1-\pi_i)\, P\left(Y_{ij}=1 \,|\theta_i, a_j, b_j\right) f\left(T_{ij};\, \lambda_j, \tau_i, \sigma_j^2\right)}, \quad \text{if } Y_{ij}=1,$$

$$\frac{\pi_i \left(1-g_j\right) h\left(T_{ij};\, \mu_c, \sigma_c^2\right)}{\pi_i \left(1-g_j\right) h\left(T_{ij};\, \mu_c, \sigma_c^2\right) + (1-\pi_i)\, P\left(Y_{ij}=0 \,|\theta_i, a_j, b_j\right) f\left(T_{ij};\, \lambda_j, \tau_i, \sigma_j^2\right)}, \quad \text{if } Y_{ij}=0.$$

Step 6: Sample π_i. Given a $Beta\,(\iota_1, \iota_2)$ prior and $\sum_{j=1}^{J} \eta_{ij} \sim Binomial\,(J, \pi_i)$, the fully conditional posterior of π_i is

$$\pi_i \sim Beta\left(\iota_1 + \sum_{j=1}^{J} \eta_{ij},\ \iota_2 + J - \sum_{j=1}^{J} \eta_{ij}\right).$$

Step 7: Sample g_j. Given a $Beta\,(\iota_3, \iota_4)$ prior, within the guessing behavior category $\eta_{ij} = 1$, the total number of people engaging in rapid guessing behavior on item j is $\sum_{i=1}^{N} \eta_{ij}$, and the number of correct items is $\sum_{i=1}^{N} \eta_{ij} Y_{ij}$; thus, $\sum_{i=1}^{N} \eta_{ij} Y_{ij} \sim Binomial\left(\sum_{i=1}^{N} \eta_{ij}, g_j\right)$. The fully conditional posterior is

$$g_j \sim Beta\left(\iota_3 + \sum_{i=1}^{N} \eta_{ij} Y_{ij},\ \iota_4 + \sum_{i=1}^{N} \eta_{ij} - \sum_{i=1}^{N} \eta_{ij} Y_{ij}\right).$$

Step 8: Sample τ_i. The conditional prior distribution of τ_i is assumed to follow a normal distribution with mean $\mu_{\tau|\theta} = \mu_\tau + \sigma_{\tau\theta}\sigma_\theta^{-2}(\theta_i - \mu_\theta)$ and $\sigma^2_{\tau|\theta} = \sigma^2_\tau - \sigma_{\tau\theta}\sigma_\theta^{-2}\sigma_{\theta\tau}$. That is, $\tau_i \sim N\left(\mu_{\tau|\theta}, \sigma^2_{\tau|\theta}\right)$. The fully conditional posterior distribution of τ_i given $\boldsymbol{T}^{obs}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \sigma_j^2, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P, \boldsymbol{\eta}$ is proportional to

$$\prod_{j=1}^{J} f\left(t_{ij};\, \lambda_{jv}, \tau_i, \sigma_j^2\right)^{(1-\eta_{ij})} p\left(\tau_i \left| \mu_{\theta|\tau}, \sigma^2_{\theta|\tau}\right.\right).$$

The fully conditional posterior distribution of τ_i is

$$N\left(\sigma^2_{\tau_i^*}\left(\frac{\sigma_{\theta\tau}\theta_i}{\sigma^2_\tau - \sigma^2_{\theta\tau}} + \sum_{j=1}^{J}\left[\left(1-\eta_{ij}\right)\sigma_j^{-2}\left(\lambda_j - \log t_{ij}\right)\right]\right),\ \sigma^2_{\tau_i^*}\right),$$

where $\sigma^2_{\tau_i^*} = \left(\left(\sigma^2_\tau - \sigma^2_{\theta\tau}\right)^{-1} + \sum_{j=1}^{J}\left[\left(1-\eta_{ij}\right)\sigma_j^{-2}\right]\right)^{-1}$.

Step 9: Sample λ_j. The fully conditional posterior distribution of the intensity parameter given the parameters $\boldsymbol{T}^{obs}$, $\boldsymbol{\tau}$, σ_j^2, $\boldsymbol{\mu}_I$, $\boldsymbol{\Sigma}_I$, $\boldsymbol{\eta}$ is

$$p\left(\lambda_j \left| \boldsymbol{T}_j^{obs}, \boldsymbol{\tau}, \sigma_j^2, \mu_\lambda, \sigma_\lambda^2, \boldsymbol{\eta}\right.\right) \propto \prod_{i=1}^{N} f\left(t_{ij};\, \lambda_j, \tau_i, \sigma_j^2\right)^{(1-\eta_{ij})} p\left(\lambda_j \left| \mu_\lambda, \sigma_\lambda^2\right.\right),$$

where $\lambda_j \sim N\left(\mu_\lambda, \sigma_\lambda^2\right)$. The fully conditional posterior distribution of λ_j is

$$N\left(\sigma^2_{\lambda_j^*}\left(\mu_\lambda \sigma_\lambda^{-2} + \sum_{i=1}^{N}\left(1-\eta_{ij}\right)\left(\log t_{ij} + \tau_i\right)\sigma_j^{-2}\right), \sigma^2_{\lambda_j^*}\right),$$

where $\sigma^2_{\lambda_j^*} = \left(\sigma_\lambda^{-2} + \sigma_j^{-2}\sum_{i=1}^{N}\left(1-\eta_{ij}\right)\right)^{-1}$.

Step 10: Sample σ_j^2. A prior for σ_j^2 is an inverse-gamma distribution, $IG\,(\upsilon_1, \omega_1)$. The fully conditional posterior distribution of σ_j^2 is

$$IG\left(\upsilon_1 + \frac{\sum_{i=1}^{N}\left(1-\eta_{ij}\right)}{2}, \omega_1 + \frac{\sum_{i=1}^{N}\left[\left(1-\eta_{ij}\right)\left(\log t_{ij} - \lambda_j + \tau_i\right)^2\right]}{2}\right).$$

Step 11: Sample μ_c. We assume a uniform prior $p\,(\mu_c) \propto 1$. The fully conditional posterior distribution of μ_c is proportional to

$$p\left(\mu_c \,|\boldsymbol{T}, \boldsymbol{\eta}\right) \propto \prod_{i=1}^{N}\prod_{j=1}^{J} f\left(t_{ij};\, \mu_c, \sigma_c^2\right)^{\eta_{ij}} p\left(\mu_c\right).$$

The fully conditional posterior distribution of μ_c is

$$\mu_c \,|\boldsymbol{T}, \boldsymbol{\eta} \sim N\left(\left(\sum_{i=1}^{N}\sum_{j=1}^{J}\eta_{ij}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{j=1}^{J}\eta_{ij}\log t_{ij}\right), \left(\sum_{i=1}^{N}\sum_{j=1}^{J}\eta_{ij}\right)^{-1}\sigma_c^2\right).$$

Step 12: Sample σ_c^2. We assume that the variance parameter follows an inverse-gamma prior distribution, $IG\,(\upsilon_2, \omega_2)$. The fully conditional posterior distribution of σ_c^2 given $\boldsymbol{T}$, $\mu_c, \upsilon_2, \omega_2, \boldsymbol{\eta}$ is proportional to

$$p\left(\sigma_c^2 \,|\boldsymbol{T}, \mu_c, \upsilon_2, \omega_2, \boldsymbol{\eta}\right) \propto \prod_{i=1}^{N}\prod_{j=1}^{J} f\left(t_{ij};\, \mu_c, \sigma_c^2\right)^{\eta_{ij}} p\left(\sigma_c^2\right).$$

The fully conditional posterior distribution of σ_c^2 is

$$\sigma_c^2 \,|\boldsymbol{T}, \mu_c, \upsilon_2, \omega_2, \boldsymbol{\eta} \sim IG\left(\upsilon_1 + \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\eta_{ij}}{2}, \omega_1 + \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\eta_{ij}\left(\log t_{ij} - \mu_c\right)^2}{2}\right).$$

3.2. Metropolis–Hastings Sampling Algorithm

In order to estimate the constrained covariance matrix $\boldsymbol{\Sigma}_P = \begin{pmatrix} 1 & \sigma_{\theta\tau} \\ \sigma_{\tau\theta} & \sigma_\tau^2 \end{pmatrix}$ (where σ_θ^2 is restricted to be equal to 1 owing to the model identification limitation), we need to update each element of the constrained covariance matrix separately using the Metropolis–Hastings algorithm.

Step 13: Sample the correlation $\sigma_{\theta\tau}$ between $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$. The identification constraints induce a restricted covariance matrix. The new value $\sigma^*_{\theta\tau}$ is sampled from a truncated normal distribution $N\left(\sigma^{(r-1)}_{\theta\tau}, s^2_{01}\right) I\left(-p_{01} < \sigma^*_{\theta\tau} < p_{01}\right)$, where $p_{01} = \sqrt{\sigma^{2,(r-1)}_\tau}$. Therefore, the probability of acceptance $\alpha\left(\sigma^{(r-1)}_{\theta\tau}, \sigma^*_{\theta\tau}\right)$ can be written as

$$\min\left\{1, \frac{\prod_{i=1}^{N} p\left(\tau_i \middle| \theta_i^{(r)}, \sigma_\tau^{2,(r-1)}, \sigma^*_{\theta\tau}\right) p\left(\sigma^*_{\theta\tau}\right)\left(\Phi\left(\frac{p_{01}-\sigma^*_{\theta\tau}}{s_{01}}\right) - \Phi\left(\frac{-p_{01}-\sigma^*_{\theta\tau}}{s_{01}}\right)\right)}{\prod_{i=1}^{N} p\left(\tau_i \middle| \theta_i^{(r)}, \sigma_\tau^{2,(r-1)}, \sigma^{(r-1)}_{\theta\tau}\right) p\left(\sigma^{(r-1)}_{\theta\tau}\right)\left(\Phi\left(\frac{p_{01}-\sigma^{(r-1)}_{\theta\tau}}{s_{01}}\right) - \Phi\left(\frac{-p_{01}-\sigma^{(r-1)}_{\theta\tau}}{s_{01}}\right)\right)}\right\};$$

otherwise, $\sigma^{(r-1)}_{\theta\tau} = \sigma^*_{\theta\tau}$, where $p(\tau_i|\theta_i)$ is the conditional density function of the speed parameter, s^2_{01} is the proposal variance, and $p(\sigma_{\theta\tau})$ is the density of the uniform prior.

Step 14: Sample σ^2_τ. The new value $\sigma^{2,*}_\tau$ is sampled from a truncated normal distribution $N\left(\sigma^{2,(r-1)}_\tau, s^2_{02}\right) I\left(\sigma^{2,*}_\tau > \left(\sigma^*_{\theta^{(2)}\tau}\right)^2 = p_0\right)$. Therefore, the probability of acceptance $\alpha\left(\sigma^{2,(r-1)}_\tau, \sigma^{2,*}_\tau\right)$ can be written as

$$\min\left\{1, \frac{\prod_{i=1}^{N} p\left(\tau_i \middle| \theta_i^{(r)}, \sigma_\tau^{2,*}, \sigma^{(r)}_{\theta\tau}\right) p\left(\sigma^{2,*}_\tau; \kappa, \vartheta\right)\left(1 - \Phi\left(\frac{p_0-\sigma^{2,(r-1)}_\tau}{s_{02}}\right)\right)}{\prod_{i=1}^{N} p\left(\tau_i \middle| \theta_i^{(r)}, \sigma_\tau^{2,(r-1)}, \sigma^{(r)}_{\theta\tau}\right) p\left(\sigma^{2,(r-1)}_\tau; \kappa, \vartheta\right)\left(1 - \Phi\left(\frac{p_0-\sigma^{2,*}_\tau}{s_{02}}\right)\right)}\right\};$$

otherwise, $\sigma^{2,*}_\tau = \sigma^{2,(r-1)}_\tau$, where s^2_{02} is the proposal variance, and $p\left(\sigma^2_\tau; \kappa, \vartheta\right)$ is the density function of the scaled inverse chi-squared distribution with degrees of freedom and the scale parameter.

3.3. Bayesian Model Assessment

Two Bayesian model assessment methods were developed to evaluate the fit of the two models. The new model is a mixture model that combines responses and response times to detect rapid guessing behavior. The other model does not consider the mixture structure. Spiegelhalter et al. (2002) proposed the deviance information criterion (DIC) as a way to evaluate model fit based on Bayesian posterior estimates by considering the trade-off relationship between the adequacy of the model fitting and the number of model parameters. Write $\boldsymbol{\Lambda} = \left(\boldsymbol{\Lambda}_{ij},\ i = 1, ..., N.\, j = 1, ..., J.\right)$, where $\boldsymbol{\Lambda}_{ij} = \left(\eta_{ij}, \theta_i, a_j, b_j, \lambda_j, \tau_i, \sigma_j^2, \mu_c, \sigma_c^2, g_j, \pi_i\right)'$. Let $\left\{\boldsymbol{\Lambda}^{(1)}, ..., \boldsymbol{\Lambda}^{(M)}\right\}$, where $\boldsymbol{\Lambda}^{(m)} = \left(\eta_{ij}^{(m)}, \theta_i^{(m)}, a_j^{(m)}, b_j^{(m)}, \lambda_j^{(m)}, \tau_i^{(m)}, \sigma_j^{2,(m)}, \mu_c^{(m)}, \sigma_c^{2,(m)}, g_j^{(m)}, \pi_i^{(m)}\right)'$ for $m = 1, ..., M$, denotes an Markov chain Monte Carlo (MCMC) sample from the posterior distribution in (1). The logarithm of the joint likelihood function evaluated at $\boldsymbol{\Lambda}^{(m)}$ is given by

$$\log f\left(\boldsymbol{Y}, \boldsymbol{T} \middle| \boldsymbol{\Lambda}^{(m)}\right) = \sum_{i=1}^{N}\sum_{j=1}^{J} \log f\left(Y_{ij}, T_{ij} \middle| \boldsymbol{\Lambda}_{ij}^{(m)}\right), \quad (4)$$

where

$$\begin{aligned} &f\left(Y_{ij}, T_{ij} \middle| \boldsymbol{\Lambda}_{ij}\right) \\ &= \left[\pi_i g_j h\left(T_{ij} \middle| \mu_c, \sigma_c^2\right)\right]^{\eta_{ij}\cdot Y_{ij}} \left[\pi_i\left(1-g_j\right) h\left(T_{ij} \middle| \mu_c, \sigma_c^2\right)\right]^{\eta_{ij}\cdot\left(1-Y_{ij}\right)} \\ &\times \left[\left(1-\pi_i\right) P\left(Y_{ij}=1 \middle| \eta_{ij}=0\right) f\left(T_{ij} \middle| \lambda_j, \tau_i, \sigma_j^2\right)\right]^{(1-\eta_{ij})\cdot Y_{ij}} \\ &\times \left[\left(1-\pi_i\right) P\left(Y_{ij}=0 \middle| \eta_{ij}=0\right) f\left(T_{ij} \middle| \lambda_j, \tau_i, \sigma_j^2\right)\right]^{(1-\eta_{ij})\cdot\left(1-Y_{ij}\right)}. \end{aligned}$$

As the log-likelihood function $\log f\left(Y_{ij}, T_{ij} \middle| \boldsymbol{\Lambda}_{ij}^{(r)}\right)$, $i = 1, ..., N.\ j = 1, ..., J$, is readily available from the R outputs, $\log f\left(\boldsymbol{Y}, \boldsymbol{T} \middle| \boldsymbol{\Lambda}^{(r)}\right)$ in (4) is easy to compute. The DIC can be calculated as follows:

$$\text{DIC} = \widehat{\text{Dev}(\boldsymbol{\Lambda})} + 2P_D = \widehat{\text{Dev}(\boldsymbol{\Lambda})} + 2\left[\overline{\text{Dev}(\boldsymbol{\Lambda})} - \widehat{\text{Dev}(\boldsymbol{\Lambda})}\right], \quad (5)$$

where

$$\overline{\text{Dev}(\boldsymbol{\Lambda})} = -\frac{2}{M}\sum_{m=1}^{M} \log f\left(\boldsymbol{Y}, \boldsymbol{T} \middle| \boldsymbol{\Lambda}^{(m)}\right) \text{ and}$$

$$\widehat{\text{Dev}(\boldsymbol{\Lambda})} = -2 \max_{1\le m\le M} \log f\left(\boldsymbol{Y}, \boldsymbol{T} \middle| \boldsymbol{\Lambda}^{(m)}\right).$$

In (5), $\overline{\text{Dev}(\boldsymbol{\Lambda})}$ is a Monte Carlo estimate of the posterior expectation of the deviance function $\text{Dev}(\boldsymbol{\Lambda}) = -2\log f(\boldsymbol{Y}, \boldsymbol{T}|\boldsymbol{\Lambda})$. $\widehat{\text{Dev}(\boldsymbol{\Lambda})}$ is an approximation of $\text{Dev}(\hat{\boldsymbol{\Lambda}})$, where $\hat{\boldsymbol{\Lambda}}$ is the posterior mode, when the prior is relatively non-informative, and $P_D = \overline{\text{Dev}(\boldsymbol{\Lambda})} - \widehat{\text{Dev}(\boldsymbol{\Lambda})}$ is the effective number of parameters. The model with a smaller DIC value fits the data better.

Another method to compare the fit of the two models is to use the logarithm of the pseudomarginal likelihood (LPML; Geisser and Eddy, 1979; Ibrahim et al., 2001) by calculating the conditional predictive ordinates (CPO) index. Next, the formulas for computing the CPO and LPML are given. Letting $U_{ij,\max} = \max_{1\le m\le M}\left\{-\log f\left(Y_{ij}, T_{ij}\left|\boldsymbol{\Lambda}_{ij}^{(m)}\right.\right)\right\}$, a Monte Carlo estimate of the CPO (Gelfand et al., 1992; Chen et al., 2000) is given by

$$\log\left(\widehat{\mathrm{CPO}_{ij}}\right) = -U_{ij,\max} - \log\left[\frac{1}{M}\sum_{m=1}^{M}\exp\left\{-\log f\left(Y_{ij}, T_{ij}\left|\boldsymbol{\Lambda}_{ij}^{(m)}\right.\right) - U_{ij,\max}\right\}\right] \quad (6)$$

Note that the maximum value adjustment used in $\log\left(\widehat{\mathrm{CPO}_{ij}}\right)$ plays an important part in numerical stabilization when computing $\exp\left\{-\log f\left(Y_{ij}, T_{ij}\left|\boldsymbol{\Lambda}_{ij}^{(m)}\right.\right) - U_{ij,\max}\right\}$ in (6). A summary statistic of the $\widehat{\mathrm{CPO}_{ij}}$ is the sum of their logarithms, which is called the LPML and given by

$$\mathrm{LPML} = \sum_{i=1}^{N}\sum_{j=1}^{J}\log\left(\widehat{\mathrm{CPO}_{ij}}\right).$$

A model with a larger LPML has a better fit to the data.

TABLE 1 | The proportions of examinees and items in the simulation study 1.

	No. of items = 20	
	No. of examinees 1,000	**No. of examinees 2,000**
Item intensity	**Proportion of examinees who can not finish a 20 item test within 40 min**	
$\lambda \sim U(-0.25, 0.25)$	14.2%	12.3%
$\lambda \sim U(0.25, 0.75)$	46.6%	40.5%
Item intensity	**Proportion of items that are answered by rapid guessing**	
$\lambda \sim U(-0.25, 0.25)$	3.31%	2.88%
$\lambda \sim U(0.25, 0.75)$	14.86%	12.59%
	No. of items = 40	
	No. of examinees 1,000	**No. of examinees 2,000**
Item intensity	**Proportion of examinees who can not finish a 40 item test within 80 min**	
$\lambda \sim U(-0.25, 0.25)$	13.4%	15.4%
$\lambda \sim U(0.25, 0.75)$	44.1%	47.9%
Item intensity	**Proportion of items that are answered by rapid guessing**	
$\lambda \sim U(-0.25, 0.25)$	3.05%	3.43%
$\lambda \sim U(0.25, 0.75)$	13.90%	14.61%

4. SIMULATION STUDIES

4.1. Simulation 1

This simulation study was conducted to evaluate the recovery performance of the Pólya–gamma Gibbs sampling algorithm under different simulation conditions.

Simulation Designs

The following conditions were manipulated: (a) test length, $J = 20$ or 40, where the 20-item test is within 40 min, and the 40-item test is within 80 min; (b) the number of examinees, $N = 1,000$ or $2,000$; and (c) the speededness level, low speededness level (LSL) or high speededness level (HSL). The speededness level is controlled by the intensity parameter λ_j. That is, a larger time intensity parameter corresponds to a longer average testing time. Fully crossing the different values of these four factors yielded eight conditions (two test lengths × two sample sizes × two speededness levels).

True Values and Prior Distributions

For the 2PL model, true values of item discrimination parameters a_j are generated from a truncated normal distribution, i.e., $a_j \sim N(0,1)\,\mathrm{I}(0,+\infty)$, $j = 1, 2, ..., J$, where the indicator function $\mathrm{I}(A)$ takes a value of 1 if A is true and 0 if A is false. The item difficulty parameters b_j are generated from a standardized normal distribution. For the RT model, the response times of the rapid guessing behavior, C_{ij}, are generated from a log-normal distribution (Wang and Xu, 2015, p. 464), i.e., $\log C_{ij} \sim N(-2, 0.25)$. The correct response probability of the rapid guessing behavior, g_j, is set to 0.25 for all items (Wang and Xu, 2015). Although the variances of the RT model, σ_j^2, can vary across items in the process of model setting and algorithm implementation, for convenience, we assume that the variance of the RT model, σ_j^2, is set to 0.5 for all items. We controlled the speededness level by adjusting the time intensity parameter, that is, low speededness distribution $\lambda \sim U(-0.25, 0.25)$ and high speededness distribution $\lambda \sim U(0.25, 0.75)$. The proportion of examinees who could not finish a test within the allocated time is shown in **Table 1**. The proportion of items that were answered by guessing is also shown in **Table 1**. For the population distribution of person parameters, the ability and speed parameters $(\theta, \tau)'$ were generated from a bivariate normal distribution with mean vector $(0, 0)'$ and covariance matrix $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.25 \end{pmatrix}$. The responses and response times were generated from the 2PL model and log-normal distribution. The following method was used to generate the guessing behavior indicator η_{ij}. For all items, examinees could finish a given test within the allotted time having $\eta_{ij} = 0$, where $j = 1, ..., J$. Other η_{ij} were generated by the following two steps. Assuming that the generated response time data has no time limit for all items, then we replace T_{ij} with C_{ij} from the last item backward until the total response time is less than or equal to the allocated time. Therefore, given the eight simulation conditions, the RT paths for the examinees are shown in **Figures 1, 2**. **Figures 3**, **4** show the histograms of response times obtained from all item–person combinations**.** The non-informative priors and hyperpriors for the parameters were chosen as follows: $p(a_j) \sim N\left(0, 10^5\right)\mathrm{I}(0,+\infty)$,

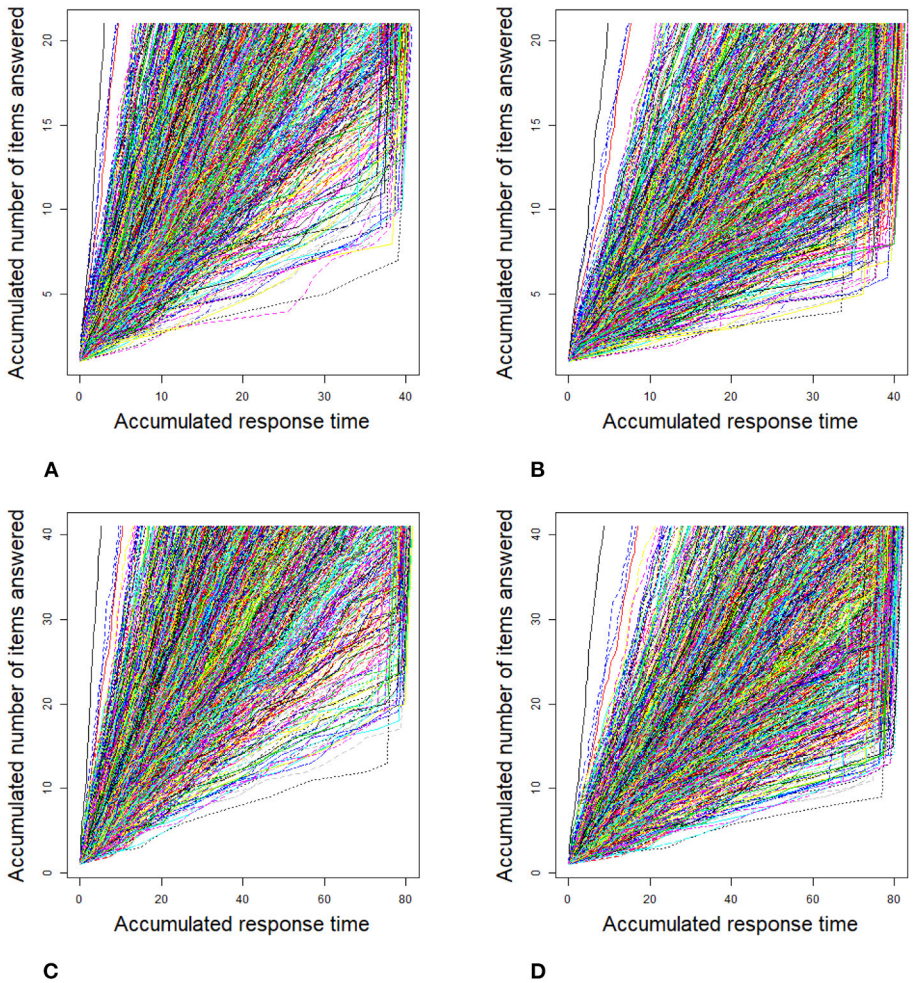

FIGURE 1 | Response time paths for 1,000 examinees at different speededness levels in the simulation study 1. **(A)** items 20 and low speededness, **(B)** items 20 and high speededness, **(C)** items 40 and low speededness, and **(D)** items 40 and high speededness.

$p\left(b_j\right) \sim N\left(0, 10^5\right)$, $p\left(g_j\right) \sim Beta\left(5, 17\right)$, $p\left(\lambda_j\right) \sim N\left(0, 10^5\right)$, $p\left(\pi_i\right) \sim Beta\left(1, 5\right)$, $p\left(\mu_c\right) \sim N\left(-3, 10^5\right)$, $p\left(\sigma_c^2\right) \sim Inv\text{-}Gamma(0.0001, 0.0001)$, $p\left(\sigma_\tau^2\right) \sim Inv\text{-}Gamma(0.0001, 0.0001)$, and $\sigma_{\theta\tau} \sim U\left(-\sqrt{\sigma_\tau^2}, \sqrt{\sigma_\tau^2}\right)$, where $\sigma_\tau^2 = 1$. Fifty replications were considered in each simulation condition.

Convergence diagnostics

In order to evaluate the convergence of parameter estimates, we only considered convergence in the case of minimum sample sizes with HSLs owing to space limitations. That is, the test length was fixed at 20, and the number of examinees was 1,000. Two methods were used to check the convergence of our algorithm: the "eyeball" method to monitor the convergence by visually inspecting the history plots of the generated sequences; and the Gelman–Rubin method (Gelman and Rubin, 1992; Brooks and Gelman, 1998).

The convergence of the Bayesian algorithm was checked by monitoring the trace plots of the parameters for consecutive sequences of 20,000 iterations. The first 10,000 iterations were set as the burn-in period. As an illustration, four chains started at overdispersed starting values were run for each replication. The trace plots of item parameters randomly selected are shown in **Figure 5**. In addition, the potential scale reduction factor (PSRF; Brooks and Gelman, 1998) values for all item parameters are shown in **Figure 6**. We found that the PSRF values for all parameters were less than 1.2, which ensured that all chains converged as expected.

5. RESULTS

As shown in **Table 2**, the bias was between 0.0098 and 0.1411 for the discrimination parameters $\boldsymbol{a}$, between −0.0335 and 0.0010 for the difficulty parameters $\boldsymbol{b}$, between −0.0206 and 0.0115 for the rapid guessing parameters $\boldsymbol{g}$, between −0.0271 and 0.0386 for the time intensity parameters $\boldsymbol{\lambda}$, between −0.0105 and 0.0314 for the time discrimination parameters $\boldsymbol{\sigma}^2$, between 0.0196 and 0.0313 for the ability parameters $\boldsymbol{\theta}$, between 0.0058 and 0.0377 for

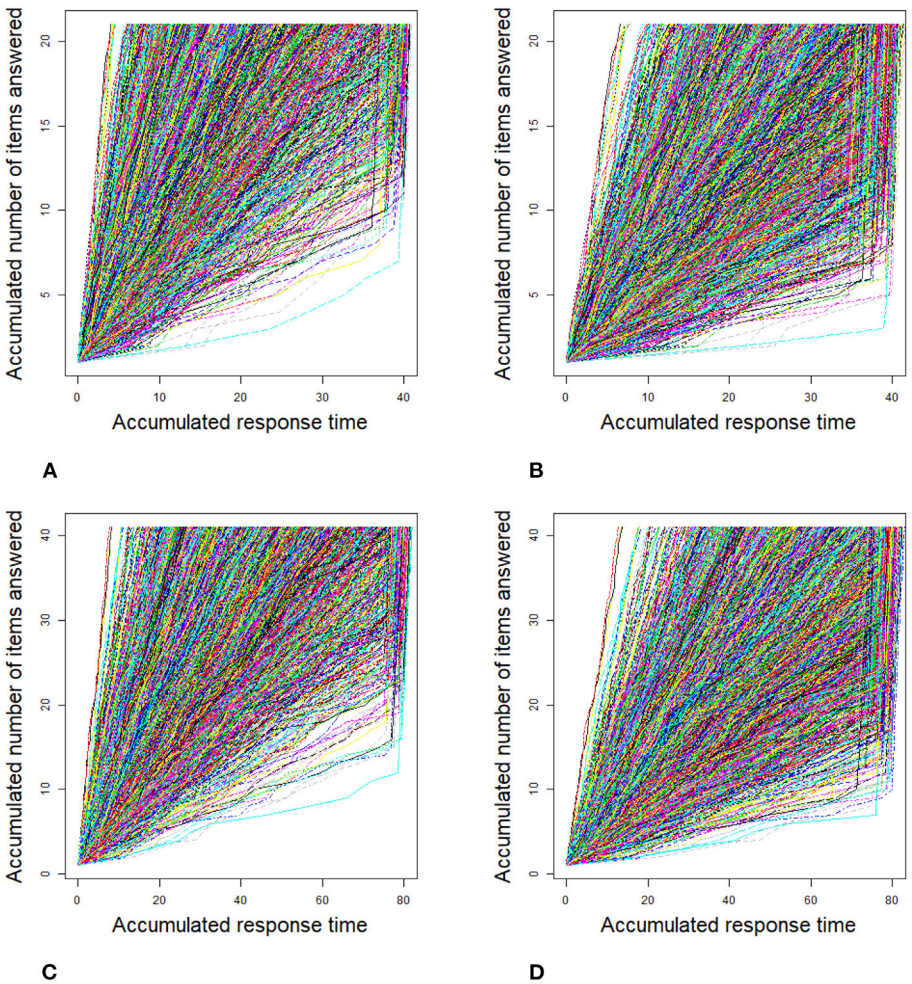

FIGURE 2 | Response time paths for 2,000 examinees at different speededness levels in the simulation study 1. **(A)** items 20 and low speededness, **(B)** items 20 and high speededness, **(C)** items 40 and low speededness, and **(D)** items 40 and high speededness.

the speed parameters $\boldsymbol{\tau}$, between −0.0259 and 0.0202 for the μ_c, between −0.0373 and 0.0136 for the σ_c^2, between −0.0671 and −0.0102 for the $\sigma_{\theta\tau}$, and between −0.0201 and 0.0056 for the σ_τ^2. In addition, the MSE was between 0.0125 and 0.0413 for the discrimination parameters $\boldsymbol{a}$, between 0.0041 and 0.0138 for the difficulty parameters $\boldsymbol{b}$, between 0.0009 and 0.0026 for the rapid guessing parameters $\boldsymbol{g}$, between 0.0001 and 0.0017 for the time intensity parameters $\boldsymbol{\lambda}$, between 0.0001 and 0.0005 for the time discrimination parameters $\boldsymbol{\sigma}^2$, between 0.0873 and 0.1920 for the ability parameters $\boldsymbol{\theta}$, between 0.0068 and 0.0693 for the speed parameters $\boldsymbol{\tau}$, between 0.0000 and 0.0007 for the μ_c, between 0.0000 and 0.0009 for the σ_c^2, between 0.0010 and 0.0045 for the $\sigma_{\theta\tau}$, and between 0.0002 and 0.0009 for the σ_τ^2. In summary, the Pólya–gamma Gibbs sampling algorithm provides accurate estimates of the parameters for various numbers of examinees and items.

5.1. Simulation 2

In this simulation study, we focus on the model fitting data for the mixture model and non-mixture model based on different simulation conditions from the perspective of Bayesian model assessment. Two Bayesian model assessment tools, DIC and LPML, are used to identify the true models.

Simulation Designs

For purposes of illustration, the numbers of examinees and items were fixed at 1,000 and 40, respectively. The true value settings for the item parameters in the 2PLIRT model and response time model were the same as in simulation study 1. The first factor is the correlation coefficient. Three correction coefficients $\rho_{\theta\tau}$ were considered in this simulation. That is, (1) $\rho_{\theta\tau} = 0.3$ (θ and τ have weak correlation; WC); (2) $\rho_{\theta\tau} = 0.8$ (θ and τ have a strong correlation; SC). Furthermore, the true values of θ and τ can be drawn from a bivariate

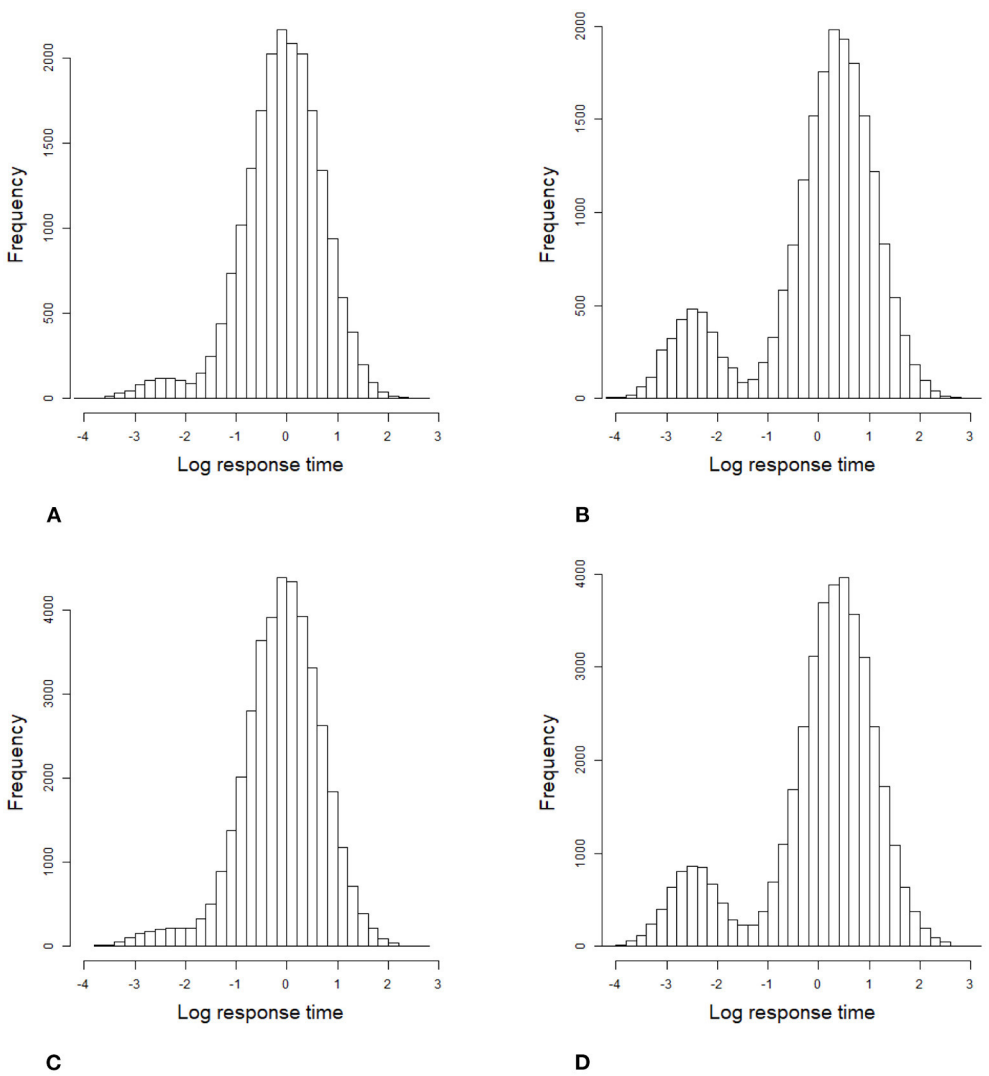

FIGURE 3 | Histogram of 1,000 examinees' response times based on all item–person combinations in the simulation study 1. **(A)** items 20 and low speededness, **(B)** items 20 and high speededness, **(C)** items 40 and low speededness, and **(D)** items 40 and high speededness.

normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\begin{pmatrix} 1 & \rho_{\theta\tau} \\ \rho_{\theta\tau} & 1 \end{pmatrix}$. The second factor is the speededness level, which was varied by adjusting the time intensity parameter λ: (1) LSL, $\lambda \sim U(-0.25, 0.25)$; (2) HSL, $\lambda \sim U(0.25, 0.75)$. The third factor is the choice of fitting model: (1) mixture model; (2) non-mixture model (hierarchical structure model of van der Linden, 2007). Based on the abovementioned test conditions, the item responses and response time data were respectively generated from the 2PLIRT model and response time model. Therefore, the true models and the fitted models were designed as follows.

(i) True model, i.e., mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$LSL **vs.** fitted model, i.e., mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$LSL, and non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$LSL.

(ii) True model, i.e., mixture model with SC ($\rho_{\theta\tau} = 0.8$)$\oplus$LSL **vs.** fitted model, i.e., mixture model with SC ($\rho_{\theta\tau} = 0.8$)$\oplus$LSL, and non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)$\oplus$LSL.

(iii) True model, i.e., mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$HSL **vs.** fitted model, i.e., mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$HSL, and non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$HSL.

(iv) True model, i.e., mixture model with SC ($\rho_{\theta\tau} = 0.8$)$\oplus$HSL **vs.** fitted model, i.e., mixture model with SC ($\rho_{\theta\tau} = 0.8$)$\oplus$HSL, and non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)$\oplus$HSL.

(v) True model, i.e., non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$LSL **vs.** fitted model, i.e., mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$LSL, and non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)$\oplus$LSL.

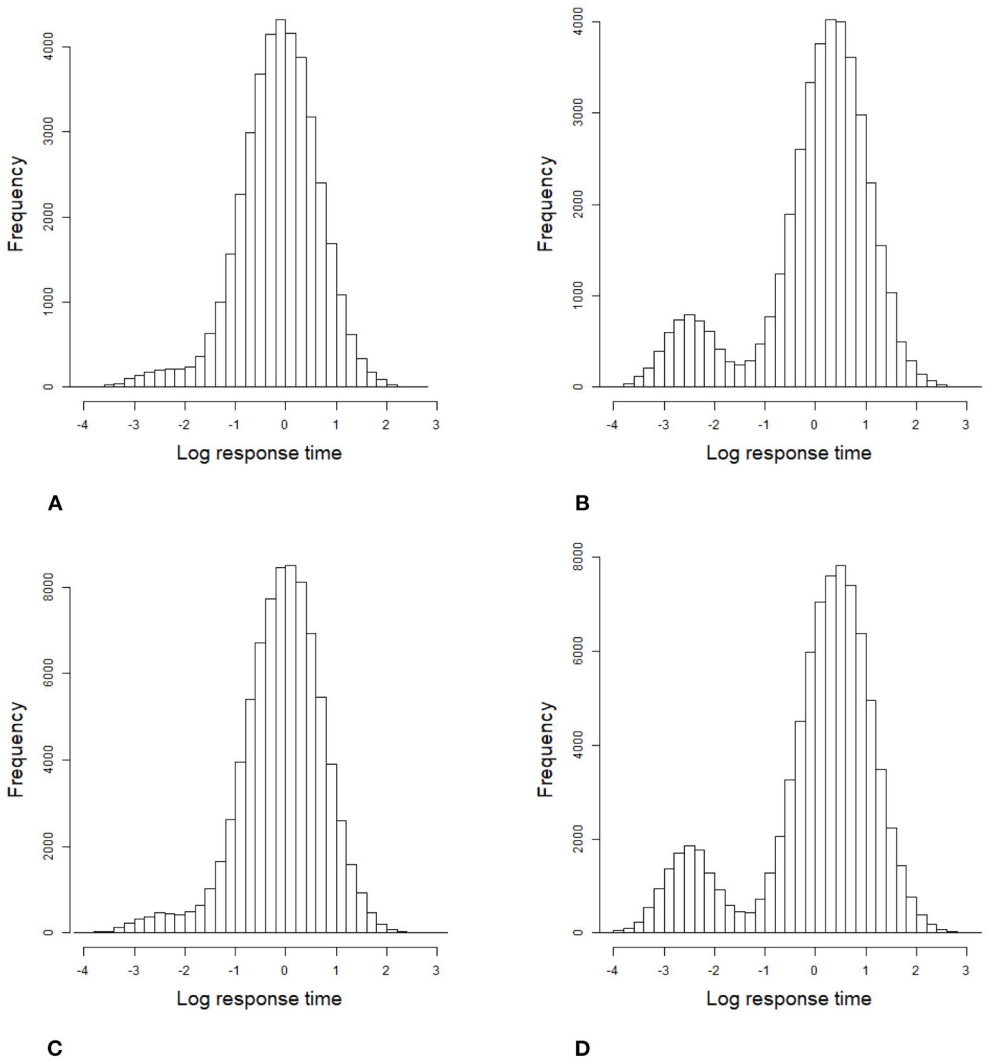

FIGURE 4 | Histogram of 2,000 examinees' response times based on all item–person combinations in the simulation study 1. **(A)** items 20 and low speededness, **(B)** items 20 and high speededness, **(C)** items 40 and low speededness, and **(D)** items 40 and high speededness.

(vi) True model, i.e., non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)⊕LSL **vs.** fitted model, i.e., mixture model with SC ($\rho_{\theta\tau} = 0.8$)⊕LSL, and non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)⊕LSL.

(vii) True model, i.e., non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)⊕HSL **vs.** fitted model, i.e., mixture model with WC ($\rho_{\theta\tau} = 0.3$)⊕HSL, and non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)⊕HSL.

(viii) True model, i.e., non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)⊕HSL **vs.** fitted model, i.e., mixture model with SC ($\rho_{\theta\tau} = 0.8$)⊕HSL, and non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)⊕HSL.

The priors of parameters were also the same as those used in simulation 1. That is, the non-informative priors were used in this simulation study. To implement the MCMC sampling algorithm, chains of length 10,000 with an initial burn-in period of 20,000 were chosen. There were 50 replications for each simulation condition. The PSRF (Brooks and Gelman, 1998) values for all item and person parameters for each simulation condition were less than 1.2.

Results

As shown in **Tables 3**, **4**, regardless of whether the speededness levels were low or high, and whether the correlation coefficients were weak ($\rho_{\theta\tau} = 0.3$) or strong ($\rho_{\theta\tau} = 0.8$), both Bayesian model assessment criteria could accurately identify the true models when the data were generated from the mixture models and non-mixture models. More specifically, under the LSL and WC conditions, when the mixture model was the true model, the mixture model fitted the data better, as expected. The median DIC of the mixture model (185007.092) was smaller than that of the non-mixture model (201335.596), and the median LPML of the mixture model (−91302.451) was larger than that of

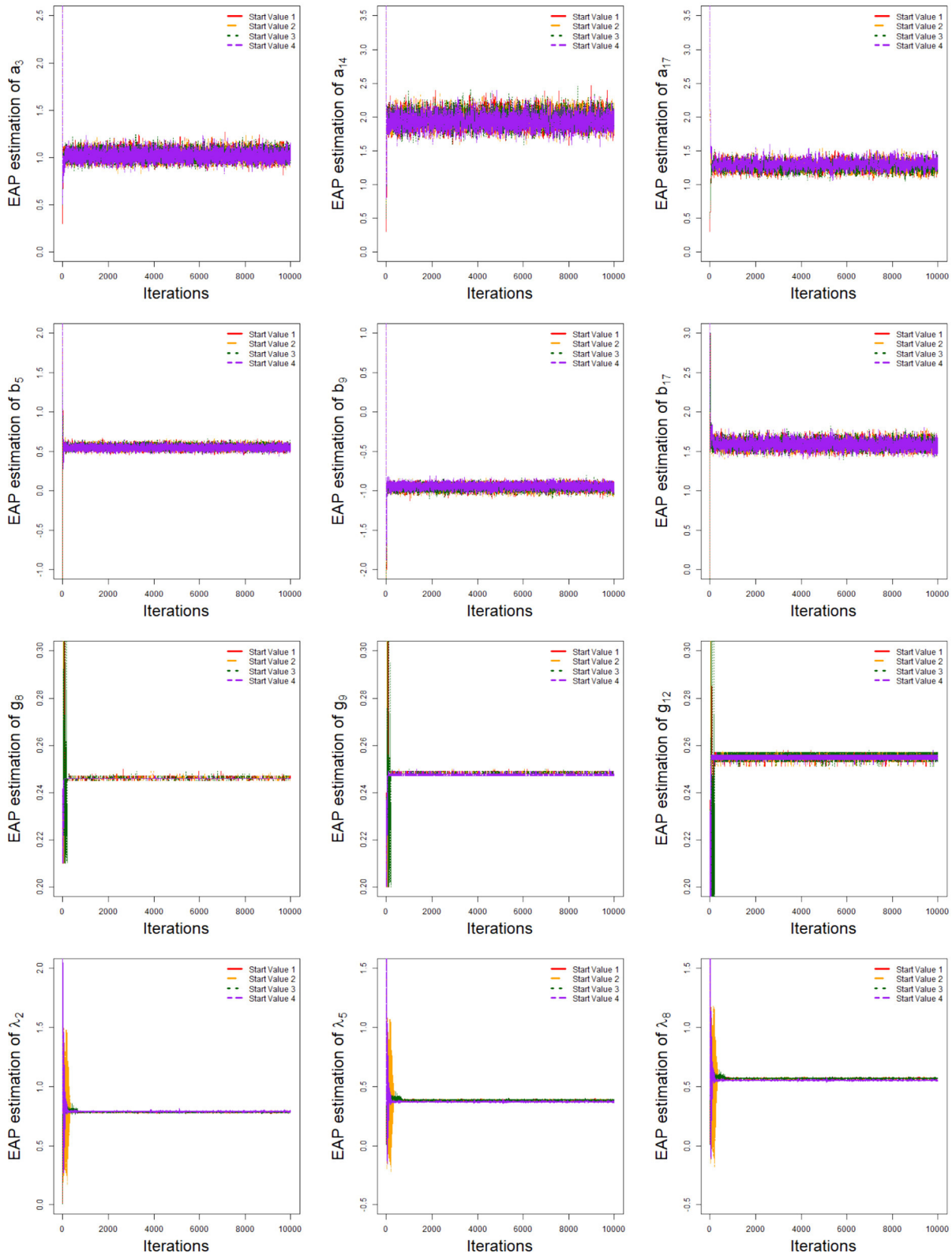

FIGURE 5 | The trace plots of three randomly selected items for the simulation study 1.

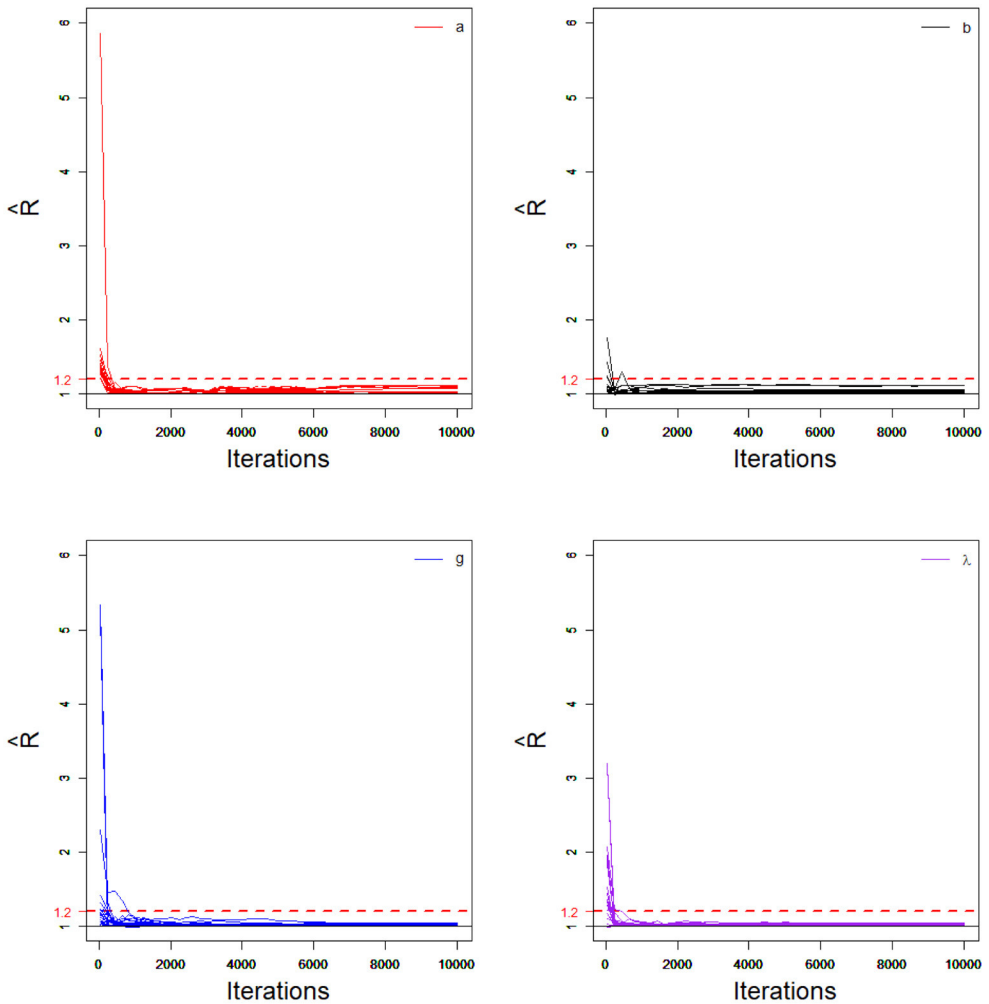

FIGURE 6 | The trace plots of $\widehat{R}$ for item parameters in the simulation study 1.

TABLE 2 | Evaluating the accuracy of parameters based on mixture model in simulation study 1.

	$N = 1,000, J = 20$		$N = 1,000, J = 40$		$N = 2,000, J = 20$		$N = 2,000, J = 40$	
	LSL	**HSL**	**LSL**	**HSL**	**LSL**	**HSL**	**LSL**	**HSL**
Bias								
a	0.0320	0.0814	0.0098	0.0291	0.1002	0.1411	0.0253	0.0472
b	−0.0149	−0.0162	−0.0252	−0.0335	−0.0203	−0.0194	0.0010	−0.0030
g	−0.0136	−0.0203	−0.0193	−0.0166	−0.0005	−0.0022	−0.0206	0.0115
λ	0.0195	−0.0169	0.0386	0.0160	0.0077	−0.0271	0.0152	−0.0100
σ^2	−0.0105	0.0058	−0.0062	−0.0041	−0.0092	0.0123	−0.0080	0.0314
θ	0.0268	0.0295	0.0210	0.0220	0.0286	0.0313	0.0196	0.0260
τ	0.0214	0.0137	0.0377	0.0218	0.0098	0.0058	0.0168	0.0152
μ_c	−0.0259	0.0092	−0.0108	0.0078	−0.0226	0.0069	0.0041	0.0202
σ_c^2	−0.0373	0.0132	−0.0371	0.0115	−0.0371	0.0063	−0.0331	0.0136
$\sigma_{\theta\tau}$	−0.0474	−0.0671	−0.0373	−0.0501	−0.0324	−0.0552	−0.0119	−0.0102
σ_τ^2	−0.0182	−0.0201	−0.0049	−0.0103	−0.0057	−0.0054	0.0037	0.0056

(Continued)

TABLE 2 | Continued

	$N = 1,000, J = 20$		$N = 1,000, J = 40$		$N = 2,000, J = 20$		$N = 2,000, J = 40$	
	LSL	**HSL**	**LSL**	**HSL**	**LSL**	**HSL**	**LSL**	**HSL**
MSE								
$\boldsymbol{a}$	0.0252	0.0355	0.0287	0.0375	0.0413	0.0527	0.0125	0.0167
$\boldsymbol{b}$	0.0071	0.0085	0.0105	0.0138	0.0084	0.0108	0.0041	0.0058
$\boldsymbol{g}$	0.0026	0.0018	0.0014	0.0011	0.0013	0.0010	0.0015	0.0009
$\boldsymbol{\lambda}$	0.0007	0.0011	0.0017	0.0006	0.0001	0.0013	0.0003	0.0015
$\boldsymbol{\sigma}^2$	0.0001	0.0004	0.0001	0.0002	0.0001	0.0005	0.0001	0.0003
$\boldsymbol{\theta}$	0.1587	0.1841	0.0943	0.1107	0.1711	0.1920	0.0873	0.1155
$\boldsymbol{\tau}$	0.0133	0.0557	0.0080	0.0141	0.0148	0.0693	0.0068	0.0099
μ_C	0.0006	0.0000	0.0001	0.0000	0.0005	0.0000	0.0000	0.0007
σ_C^2	0.0003	0.0001	0.0006	0.0001	0.0005	0.0000	0.0009	0.0001
$\sigma_{\theta\tau}$	0.0022	0.0045	0.0014	0.0025	0.0010	0.0030	0.0015	0.0026
σ_τ^2	0.0003	0.0004	0.0002	0.0005	0.0002	0.0006	0.0005	0.0009

Note that the Bias and MSE denote the average Bias and MSE for the interested parameters. $\boldsymbol{a}$ *represents all discrimination parameters,* $\boldsymbol{b}$ *represents all difficulty parameters,* $\boldsymbol{g}$ *represents all rapid guessing parameters,* $\boldsymbol{\lambda}$ *represents all time intensity parameters,* $\boldsymbol{\sigma}^2$ *represents all time discrimination parameters,* $\boldsymbol{\theta}$ *represents all ability parameters, and* $\boldsymbol{\tau}$ *represents all speed parameters.*

TABLE 3 | The results of Bayesian model assessment in simulation study 2.

Low speededness level (LSL)					
	Fitted model			**Mixture model with WC**	**Non-mixture model with WC**
True model	Mixture model with WC ($\rho_{\theta\tau} = 0.3$)	DIC	Q_1	183970.082	200906.367
			Median	185007.092	201335.596
			Q_3	185472.819	201700.856
		LPML	Q_1	–91433.366	–103949.160
			Median	–91302.451	–103871.796
			Q_3	–91095.166	–103782.198
Low speededness level (LSL)					
	Fitted model			**Mixture model with SC**	**Non-mixture model with SC**
True Model	Mixture model with SC ($\rho_{\theta\tau} = 0.8$)	DIC	Q_1	182423.016	200490.494
			Median	182806.907	200960.661
			Q_3	183285.554	201204.742
		LPML	Q_1	–91270.116	–103687.867
			Median	–91213.797	–103584.228
			Q_3	–91100.563	–103419.208
High speededness level (HSL)					
	Fitted model			**Mixture model with WC**	**Non-mixture model with WC**
True Model	Mixture model with WC ($\rho_{\theta\tau} = 0.3$)	DIC	Q_1	159487.663	175985.981
			Median	159985.584	176499.862
			Q_3	161227.782	176989.732
		LPML	Q_1	–80685.663	–87906.257
			Median	–80474.893	-87782.508
			Q_3	–80332.172	–87673.533
High speededness level (HSL)					
	Fitted model			**Mixture model with SC**	**Non-mixture model with SC**
True Model	Mixture model with SC ($\rho_{\theta\tau} = 0.8$)	DIC	Q_1	159235.762	175815.800
			Median	159629.846	176335.113
			Q_3	160570.239	176859.457
		LPML	Q_1	–80840.626	–87917.891
			Median	–80736.678	–87714.244
			Q_3	–80570.342	–87638.130

Note that the mixture model is the model in Section 2. The non-mixture model is the hierarchical structure model in van der Linden (2007).

TABLE 4 | The results of Bayesian model assessment in simulation study 2.

Low speededness level (LSL)					
		Fitted model		**Mixture model with WC**	**Non-mixture model with WC**
True Model	Non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)	DIC	Q_1	191642.341	187822.030
			Median	192051.824	188057.725
			Q_3	192465.323	188289.444
		LPML	Q_1	–95287.618	–93306.447
			Median	–95204.235	–93222.498
			Q_3	–95146.751	–93168.033
Low speededness level (LSL)					
		Fitted model		**Mixture model with SC**	**Non-mixture model with SC**
True Model	Non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)	DIC	Q_1	191663.580	187582.329
			Median	192059.746	187868.073
			Q_3	192341.397	187988.874
		LPML	Q_1	–95293.492	–93319.285
			Median	–95177.928	-93224.461
			Q_3	–95127.793	–93132.479
High speededness level (HSL)					
		model	Fitted	**Mixture model with WC**	**Non-mixture model with WC**
True Model	Non-mixture model with WC ($\rho_{\theta\tau} = 0.3$)	DIC	Q_1	191880.178	187523.642
			Median	192161.323	187831.945
			Q_3	192528.860	188102.832
		LPML	Q_1	–95194.438	–93202.085
			Median	–95108.402	-93129.144
			Q_3	–94999.978	–93038.260
High speededness level (HSL)					
		Fitted model		**Mixture model with SC**	**Non-mixture model with SC**
True Model	Non-mixture model with SC ($\rho_{\theta\tau} = 0.8$)	DIC	Q_1	191396.999	187321.113
			Median	191702.770	187686.570
			Q_3	192171.363	187941.382
		LPML	Q_1	–95202.124	–93221.728
			Median	–95101.015	–93157.626
			Q_3	–95012.373	–93028.099

Note that the mixture model is the model in Section 2. The non-mixture model is the hierarchical structure model in van der Linden (2007).

the non-mixture model (−103871.796). Similarly, under the HSL and SC conditions, when the mixture model was the true model, the mixture model also fitted the data best. The differences in the medians of DIC and LPML between the mixture model and non-mixture model were −16705.267 and 6977.566, respectively. In addition, under the LSL and WC conditions, when the non-mixture model was the true model, the non-mixture model fitted the data better. The median DIC of the non-mixture model (188057.725) was smaller than that of the mixture model (192051.824), and the median LPML of the non-mixture model (−93222.498) was larger than that of the mixture model (−95204.235). Similarly, under the HSL and SC conditions, when

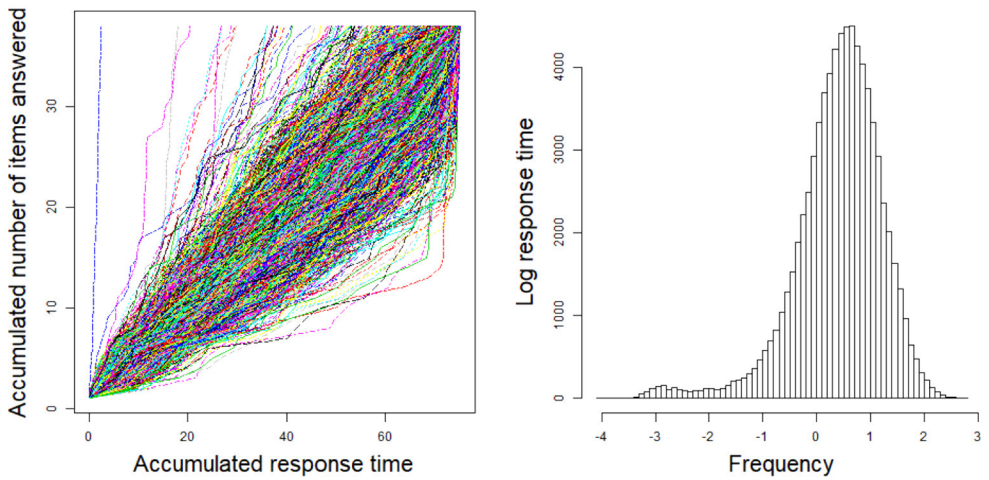

FIGURE 7 | The response time path for the examinees and the histogram of response times obtained from all item-person combinations.

TABLE 5 | The estimation results of discrimination and difficulty parameter for the real data.

Para.		EAP		SD		HPDI	
a	b	$\hat{a}$	$\hat{b}$	SD_a	SD_b	$HPDI_a$	$HPDI_b$
a_1	b_1	0.8182	−0.8649	0.0009	0.0008	[0.7591, 0.8832]	[−0.9234, −0.8073]
a_2	b_2	0.7302	−1.1924	0.0007	0.0015	[0.6809, 0.7862]	[−1.2680, −1.1134]
a_3	b_3	0.4409	−1.2129	0.0003	0.0028	[0.4034, 0.4786]	[−1.3152, −1.1096]
a_4	b_4	0.2000	−1.0279	0.0000	0.0030	[0.1863, 0.2036]	[−1.1353, −0.9183]
a_5	b_5	0.6192	−0.7536	0.0007	0.0010	[0.5652, 0.6715]	[−0.8159, −0.6888]
a_6	b_6	0.5618	−1.0134	0.0005	0.0016	[0.5150, 0.6075]	[−1.0982, −0.9389]
a_7	b_7	0.6946	−1.8531	0.0005	0.0027	[0.6518, 0.7405]	[−1.9656, −1.7591]
a_8	b_8	0.3710	−1.3215	0.0003	0.0042	[0.3350, 0.4046]	[−1.4438, −1.1925]
a_9	b_9	0.5969	−0.6650	0.0008	0.0010	[0.5441, 0.6552]	[−0.7280, −0.6072]
a_{10}	b_{10}	0.6228	−0.9849	0.0007	0.0015	[0.5738, 0.6769]	[−1.0609, −0.9129]
a_{11}	b_{11}	0.5124	−0.1673	0.0008	0.0004	[0.4601, 0.5719]	[−0.2073, −0.1293]
a_{12}	b_{12}	0.7251	−0.8260	0.0008	0.0009	[0.6674, 0.7812]	[−0.8851, −0.7662]
a_{13}	b_{13}	0.3342	−1.5034	0.0002	0.0058	[0.3011, 0.3663]	[−1.6613, −1.3594]
a_{14}	b_{14}	0.5786	−0.0406	0.0008	0.0003	[0.5179, 0.6319]	[−0.0784, −0.0093]
a_{15}	b_{15}	0.3464	−1.2434	0.0003	0.0045	[0.3140, 0.3846]	[−1.3769, −1.1141]
a_{16}	b_{16}	1.0816	−0.8625	0.0006	0.0006	[1.0050, 1.1628]	[−0.9109, −0.8092]
a_{17}	b_{17}	0.4434	−1.3966	0.0003	0.0035	[0.4070, 0.4823]	[−1.5151, −1.2828]
a_{18}	b_{18}	0.6631	−0.2462	0.0010	0.0003	[0.6023, 0.7263]	[−0.2826, −0.2071]
a_{19}	b_{19}	0.5072	−0.8406	0.0005	0.0015	[0.4600, 0.5525]	[−0.9186, −0.7620]
a_{20}	b_{20}	0.2638	−0.7837	0.0003	0.0042	[0.2251, 0.2972]	[−0.9173, −0.6637]
a_{21}	b_{21}	0.5548	−0.7497	0.0006	0.0012	[0.5030, 0.6056]	[−0.8212, −0.6832]
a_{22}	b_{22}	0.6791	−0.4723	0.0010	0.0006	[0.6150, 0.7403]	[−0.5235, −0.4273]
a_{23}	b_{23}	0.4225	−0.7727	0.0005	0.0019	[0.3803, 0.4670]	[−0.8579, −0.6881]
a_{24}	b_{24}	0.7590	−0.5959	0.0011	0.0006	[0.6925, 0.8225]	[−0.6477, −0.5447]
a_{25}	b_{25}	0.8798	−0.6894	0.0012	0.0006	[0.8136, 0.9525]	[−0.7414, −0.6393]
a_{26}	b_{26}	0.7344	−0.4227	0.0011	0.0005	[0.6680, 0.7990]	[−0.4683, −0.3774]
a_{27}	b_{27}	0.5176	−0.6252	0.0007	0.0013	[0.4685, 0.5720]	[−0.6943, −0.5492]
a_{28}	b_{28}	0.7185	−0.7225	0.0009	0.0009	[0.6601, 0.7822]	[−0.7846, −0.6619]

(Continued)

TABLE 5 | Continued

Para.		EAP		SD		HPDI	
a	b	$\widehat{a}$	$\widehat{b}$	SD_a	SD_b	$HPDI_a$	$HPDI_b$
a_{29}	b_{29}	0.7444	−0.7613	0.0009	0.0009	[0.6797, 0.8024]	[−0.8245, −0.7029]
a_{30}	b_{30}	0.5110	−0.4083	0.0007	0.0008	[0.4550, 0.5658]	[−0.4709, −0.3542]
a_{31}	b_{31}	0.4307	−0.0292	0.0007	0.0009	[0.3775, 0.4843]	[−0.0911, 0.0303]
a_{32}	b_{32}	0.7277	−0.4895	0.0011	0.0008	[0.6624, 0.7954]	[−0.5451, −0.4327]
a_{33}	b_{33}	0.5667	0.0485	0.0009	0.0004	[0.5097, 0.6253]	[0.0035, 0.0905]
a_{34}	b_{34}	0.2024	−0.7727	0.0000	0.0067	[0.2000, 0.2152]	[−0.9325, −0.6074]
a_{35}	b_{35}	0.6925	−0.6144	0.0012	0.0029	[0.6239, 0.7624]	[−0.7182, −0.5086]
a_{36}	b_{36}	0.6983	−0.2498	0.0014	0.0064	[0.6228, 0.7744]	[−0.3874, −0.0890]
a_{37}	b_{37}	0.4374	0.4227	0.0017	0.0097	[0.3525, 0.5189]	[0.1555, 0.6958]

Para. denotes the interest parameters. EAP denotes the expected a priori estimation. SD denotes the standard deviation. HPDI denotes the 95% highest posterior density intervals.

the non-mixture model was the true model, the mixture model also fitted the data better. The differences in the medians of DIC and LPML between the non-mixture model and mixture model were −4016.200 and 1943.389, respectively. Refer to **Tables 3**, **4** for more detailed results of the model assessment. In summary, the Bayesian assessment criteria were effective for identifying the true models and could, thus, be used in the subsequent real data study.

6. EMPIRICAL EXAMPLE

This section presents an application of the mixture model with an empirical example. The data set was from a high-state, large-scale, standardized computerized adaptive test that was previously analyzed by Wang and Xu (2015). The data set included 37 dichotomous items, and the test time was 75 min. The sample size was 2,106. The mixture model and non-mixture model were used to fit the item response and response time data of the 37 dichotomous items. The response time path for the examinees is shown in **Figure 7**. In addition, **Figure 7** shows a histogram of response times obtained from all item–person combinations.

In the Bayesian computation, we used 20,000 MCMC samples after a burn-in of 10,000 iterations to compute all posterior estimates. The convergence of the chains was checked using the PSRF. The PSRF values of all item parameters were less than 1.2. We used the DIC and LPML to fit the mixture model and non-mixture model. The mixture model resulted in a smaller DIC value (350696.11) than the non-mixture model (365690.66), and the LPML of the mixture model (−175027.99) was larger than that of the non-mixture model (−181062.48). This indicates that the mixture model better fitted the data. Based on the results of the model assessment, we used the mixture model to analyze real data in detail.

Analysis of item parameters

The estimated results for the discrimination and difficulty parameters are shown in **Table 5**. As shown in the table, the expected a posteriori (EAP) estimates of the one-item discrimination parameters were greater than 1. This indicated that the items could well distinguish the differences between abilities. The three items with the lowest discrimination were items 4, 34, and 20. The EAP estimates of discrimination parameters for these three items were 0.2000, 0.2024, and 0.2638. In addition, another three items had the lowest EAP estimates of the difficulty parameters, indicating that these items were easier than the other items. These were items 7, 13, and 8. The EAP estimates of g_j had a range of 0.1334 to 0.2945. The EAP estimates of λ_j had a range of −0.3322 to 0.7634.

7. CONCLUSION

In this article, we propose a novel and efficient Bayesian algorithm (Pólya–gamma Gibbs sampling algorithm) based on the auxiliary variables for estimating the mixture hierarchical model. The new algorithm avoids the tedious multidimensional integral operation of the MMLE. Within a fully Bayesian framework, the Pólya–gamma Gibbs sampling algorithm not only avoids the heavy reliance of the traditional Metropolis–Hastings algorithm on the tuning parameters of the proposed distributions for different data sets but also overcomes the disadvantage of the Metropolis–Hastings algorithm being sensitive to step size. However, the computational burden of the Pólya–gamma Gibbs sampling algorithm becomes excessive especially when there are a large number of examinees, the items or the abnormal response and response time data are considered, or a large number MCMC sample size is used. Therefore, it would be desirable to develop a stand-alone R package associated with Fortran software for a more extensive large-scale assessment program.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

ZZ completed the writing of the article. JZ and JL provided original thoughts. ZZ and JL provided key technical support. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (grant no. 12001091), China Postdoctoral Science Foundations (grant nos. 2021M690587 and 2021T140108), the Fundamental Research Funds for the Central Universities of China (grant no. 2412020QD025), Yili Normal University 2021 Annual Research Project (grant no. 2021YSBS012).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.841372/full#supplementary-material

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibb sampling. *J. Educ. Stat.* 17, 251–269.

Albert, J. H., and S. Chib. (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* 88, 669–679.

Asparouhov, T., and Muthén, B. (2010b). *Bayesian analysis of latent variable models using Mplus* (Technical report, Version 4). Available online at: http://www.statmodel.com.

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

Béguin, A. A., and Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika* 66, 541–561. doi: 10.1007/BF02296195

Biane, P., Pitman, J., and Yor, M. (2001). Probability laws related to the Jacobi theta and Riemann zeta functions, and brownian excursions. *Bull. Am. Math. Soc.* 38, 435–465. doi: 10.48550/arXiv.math/9912170

Birnbaum, A. (1968). "Some latent trait models and their use in inferring an examinee's ability," in *Statistical Theories of Mental Test Scores*, eds F. M. Lord and M. R. Novick (Reading, MA: Addison-Wesley), 397–472.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika* 46, 443–459.

Bock, R. D., and Schilling, S. G. (1997). "High-dimensional full-information item factor analysis," in *Latent Variable Modelling and Applications to Causality*, ed M. Berkane (New York, NY: Springer), 164–176.

Bolt, D. M., Cohen, A. S., and Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: application of a mixture Rasch model with ordinal constraints. *J. Educ. Meas.* 39, 331–348. doi: 10.1111/j.1745-3984.2002.tb01146.x

Boughton, K. A., and Yamamoto, K. (2007). "A HYBRID model for test speededness," in *Multivariate and Mixture Distribution Rasch Models*, eds M. von Davier and C. H. Carstensen (New York, NY: Springer), 147–156.

Brooks, S. P., and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434-455.

Chang, Y. W., Tsai, R. C., and Hsu, N. J. (2014). A speeded item response model: leave the harder till later. *Psychometrika* 79, 255–274. doi: 10.1007/s11336-013-9336-2

Chen, M.-H., Shao, Q.-M., and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York, NY: Springer.

Chib, S., and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Am. Stat.* 49, 327–335.

Converse, G., Curi, M., Oliveira, S., and Templin, J. (2021). Estimation of multidimensional item response theory models with correlated latent variables using variational autoencoders. *Mach. Learn.* 110, 1463–1480. doi: 10.1007/s10994-021-06005-7

Fox, J.-P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer.

Fox, J.-P., and Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66, 269–286. doi: 10.1007/BF02294839

Fox, J. P. (2005). Multilevel IRT using dichotomous and polytomous items. *Br. J. Math. Stat. Psychol.* 58, 145–172. doi: 10.1348/000711005X38951

Geisser, S., and Eddy, W. F. (1979). A predictive approach to model selection. *J. Am. Stat. Assoc.* 74, 153–160.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). "Model determination using predictive distributions with implementation *via* sampling-based methods (with discussion)," in *Bayesian statistics 4*, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Oxford, UK: Oxford University Press), 147–167.

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

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statisti Sci*, 7, 457–472. doi: 10.1214/ss/1177011136

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

Ghosh, M., A. Ghosh, M. Chen, and A. Agresti. (2000). Noninformative priors for one parameter item response models. *Journal of Statistical Planning and Inference*, *88*, 99-115.

Goegebeur, Y., De Boeck, P., Wollack, J. A., and Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika* 73, 65–87. doi: 10.1007/s11336-007-9031-2

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57, 97–109.

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

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. Chichester, UK: John Wiley & Sons.

Jiang, Z., and Templin, J. (2019). Gibbs samplers for logistic item response models *via* the pólya–gamma distribution: a computationally efficient data–augmentation strategy. *Psychometrika* 84, 358?374. doi: 10.1007/s11336-018-9641-x

Kuk, A. Y. C. (1999). Laplace importance sampling for generalized linear mixed models. *J. Stat. Comput. Simulat.* 63, 143–158.

Lee, S.-Y., and Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behav. Res.* 39, 653–686. doi: 10.1207/s15327906mbr3904_4

Lord, F. M., and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

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

Man, K., and Harring, J. R. (2021). Assessing preknowledge cheating *via* innovative measures: A multiple-group analysis of jointly modeling item responses, response times, and visual fixation counts. *Educ. Psychol. Meas.* 81, 441–465. doi: 10.1177/0013164420968630

Man, K., Harring, J. R., Ouyang, Y., and Thomas, S. L. (2018). Response time based nonparametric Kullback-Leibler divergence measure for detecting aberrant test-taking behavior. *Int. J. Testing* 18, 155–177. doi: 10.1080/15305058.2018.1429446

Maris, E. (1993). Additive and multiplicative models for gamma distributed variables, and their application as psychometric models for response times. *Psychometrika* 58, 445–469.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chemi. Phys.* 21, 1087–1092.

Muthén, B. O. (2010). *Bayesian Analysis in Mplus: A Brief Introduction (Incomplete Draft,Version 3)*. Available online at: http://www.statmodel.com/download/IntroBayesVersion%203.pdf.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Am. Stat. Assoc.* 108, 1339–1349. doi: 10.1080/01621459.2013.829001

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of general ized linear mixed models using adaptive quadrature. *Stata J.* 2, 1–21. doi: 10.1177/1536867X0200200101

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *J. Econometr.* 128, 301–323. doi: 10.1016/j.jeconom.2004.08.017

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., and Heathcote, A. (2015). The lognormal race: a cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika* 80, 491–513. doi: 10.1007/s11336-013-9396-3

Rouder, J. N., Sun, D., Speckman, P. L., Lu, J., and Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika* 68, 589–606. doi: 10.1007/BF02295614

Scheiblechner, H. (1979). Specific objective stochastic latency mechanisms. *J. Math. Psychol.* 19, 18–38.

Schnipke, D. L., and Scrams, D. J. (1997). Modeling response times with a two-state mixture model: a new method of measuring speededness. *J. Educ. Meas.* 34, 213–232.

Skaug, H. J. (2002). Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. *J. Comput. Graphical Stat.* 11, 458–470. doi: 10.1198/106186002760180617

Song, X.-Y., and Lee, S.-Y. (2012). A tutorial on the Bayesian approach for analyzing structural equation models. *J. Math. Psychol.* 56, 135–148. doi: 10.1016/j.jmp.2012.02.001

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

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussions). *Ann. Stat.* 22, 1701–1762.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *J. Educ. Behav. Stat.* 31, 181–204. doi: 10.3102/10769986031002181

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika* 72, 287–308. doi: 10.1007/s11336-006-1478-z

van der Linden, W. J., and Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika* 73, 365–384. doi: 10.1007/s11336-007-9046-8

Wang, C., Fan, Z., Chang, H.-H., and Douglas, J. (2013). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *J. Educ. Behav. Stat.* 38, 381–417. doi: 10.3102/1076998612461831

Wang, C., and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *Br. J. Math. Stat. Psychol.* 68, 456–477. doi: 10.1111/bmsp.12054

Wang, C., Xu, G., and Shang, Z. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika* 83, 223–254. doi: 10.1007/s11336-016-9525-x

Wise, S. L., and DeMars, C. E. (2006). An application of item response time: the effort-moderated IRT model. *J. Educ. Meas.* 43, 19–38. doi: 10.1111/j.1745-3984.2006.00002.x

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

Zhang, J. W., Lu, J., Chen, F., and Tao, J. (2019). Exploring the correlation between multiple latent variable and covariates in hierarchical data based on the multilevel multidimensional IRT model. *Front Psychol*, 10:2387. doi: 10.3389/fpsyg.2019.02387

Zhang, Z., Y., Zhang, J. W., Lu, J., and Tao, J. (2020). Bayesian estimation of the DINA model with Pólya-Gamma Gibbs algorithm. *Front. Psychol.* 11, 384. doi: 10.3389/fpsyg.2020.00384

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 | Frontiers in Psychology

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/fpsyg.2022.883137

Money Does Not Always Buy Happiness, but Are Richer People Less Happy in Their Daily Lives? It Depends on How You Analyze Income

Laura Kudrna[1]* *and Kostadin Kushlev*[2]

[1]Institute of Applied Health Research, University of Birmingham, Birmingham, United Kingdom, [2]Department of Psychology, Georgetown University, Washington, DC, United States

Do people who have more money feel happier during their daily activities? Some prior research has found no relationship between income and daily happiness when treating income as a continuous variable in OLS regressions, although results differ between studies. We re-analyzed existing data from the United States and Germany, treating household income as a categorical variable and using lowess and spline regressions to explore nonlinearities. Our analyses reveal that these methodological decisions change the results and conclusions about the relationship between income and happiness. In American and German diary data from 2010 to 2015, results for the continuous treatment of income showed a null relationship with happiness, whereas the categorization of income showed that some of those with higher incomes reported feeling less happy than some of those with lower incomes. Lowess and spline regressions suggested null results overall, and there was no evidence of a relationship between income and happiness in Experience Sampling Methodology (ESM) data. Not all analytic approaches generate the same results, which may contribute to explaining discrepant results in existing studies about the correlates of happiness. Future research should be explicit about their approaches to measuring and analyzing income when studying its relationship with subjective well-being, ideally testing different approaches, and making conclusions based on the pattern of results across approaches.

Keywords: happiness, measurement, time use, income, methodology

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Leomarich Casinillo,
Visayas State University, Philippines
Monica Violeta Achim,
Babeș-Bolyai University, Romania
Stefano Zamagni,
University of Bologna, Italy

***Correspondence:**
Laura Kudrna
l.kudrna@bham.ac.uk

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

Received: *24 February 2022*
Accepted: *08 April 2022*
Published: *31 May 2022*

Citation:
Kudrna L and Kushlev K (2022) Money Does Not Always Buy Happiness, but Are Richer People Less Happy in Their Daily Lives? It Depends on How You Analyze Income. Front. Psychol. 13:883137. doi: 10.3389/fpsyg.2022.883137

INTRODUCTION

Does having more money make someone feel happier? The answer to this longstanding question has implications for how individuals live their lives and societies are structured. It is often assumed that more income brings more happiness (with happiness broadly defined herein as hedonic feelings, while recognizing closely related constructs, including satisfaction and eudaimonia; Tiberius, 2006; Angner, 2010; Dolan and Kudrna, 2016; Sunstein, 2021). In many aspects of policy, upward income mobility is encouraged, and poverty can result in exclusion, stigmatization,

and discrimination by institutions and members of the public. More income provides people with opportunities and, sometimes, capabilities to consume more and thus satisfy more of their preferences, meet their desires and obtain more of what they want and need (Harsanyi, 1997; Sen, 1999; Nussbaum, 2008). These are all reasons to assume that higher income will bring greater happiness—or, at least, that low income will bring low happiness.

Some research challenges the assumption that earning more should lead to greater happiness. First, because people expect that more money should make them happier, people may feel less happy when their high expectations are not met (Graham and Pettinato, 2002; Nickerson et al., 2003) and they may adapt more quickly to more income than they expect (Aknin et al., 2009; Di Tella et al., 2010). Second, since the 1980s in many developed countries, the well-educated have had less leisure time than those who are not (Aguiar and Hurst, 2007) and people living in high-earning and well-educated households report feeling more time stress and dissatisfaction with their leisure time (Hamermesh and Lee, 2007; Nikolaev, 2018). The quantity of leisure time is not linearly related to happiness, with both too much and too little having a negative association (Sharif et al., 2021). Evidence also shows that people with higher incomes spend more time alone (Bianchi and Vohs, 2016). The lower quality and quantity of leisure and social time of people with higher incomes may, in turn, negatively impact their happiness, especially given there are strong links between social capital or "relational goods" and well-being (Helliwell and Putnam, 2004; Becchetti et al., 2008).

At the same time, some—but not all—evidence suggests that working class individuals tend to be more generous and empathetic than more affluent individuals (Kraus et al., 2010; Piff et al., 2010; Balakrishnan et al., 2017; Macchia and Whillans, 2022), and such kindness toward others has been associated with higher well-being (Dunn et al., 2008; Aknin et al., 2012). Relatedly, psychological research suggests that people with lower socioeconomic status have a more interdependent sense of self (Snibbe and Markus, 2005; Stephens et al., 2007). It is, therefore, possible that people high in income have lower well-being because they experience less of the internal "warm glow" (Andreoni, 1990) benefit that comes along with valuing social relationships and group membership. In theory, therefore, there are reasons to suppose that high income has both benefits and costs for well-being, and empirical evidence can inform the debate about when and whether these different perspectives are supported.

Empirical Evidence on Income and Happiness

The standard finding in existing literature is that higher income predicts greater happiness, but with a declining marginal utility (Dolan et al., 2008; Layard et al., 2008): that is, higher income is most closely associated with happiness among those with the least income and is least closely associated with happiness for those with the most income. Recently, this finding has been qualified by studies showing that the relationship between income and happiness depends on how happiness is conceptualized and measured: as an overall evaluation of one's life or as daily emotional states (Kahneman and Deaton, 2010; Killingsworth, 2021). In this vein, authors Kushlev et al. (2015) found no relationship between income and daily happiness in the American Time Use Survey (ATUS), which has recently been found for other happiness measures, too (Casinillo et al., 2020, 2021) The finding from Kushlev et al. (2015) was replicated in the German Socioeconomic Panel Survey (GSEOP) by Hudson et al. (2016), and in another analysis of the ATUS by Stone et al. (2018).

Some research has focused specifically on the effect of high income on happiness. Kahneman and Deaton (2010) conducted regression analyses using a Gallup sample of United States residents, finding that annual income beyond ~$75K was not associated with any higher daily emotional well-being. Income beyond ~$75K, however, predicted better life evaluations. Using a self-selecting sample of experiential data in the United States, Killingsworth (2021) conducted piecewise regressions and found no evidence of satiation or turning points. Jebb et al. (2018) fit regression spline models to global Gallup data, showing that the satiation point in daily experiences found by Kahneman and Deaton (2010) was also apparent in other countries. Unlike Kahneman and Deaton (2010), however, Jebb et al. (2018) also found evidence of satiation in people's life evaluations, and even some evidence for "turning points"—whereby richer people evaluated their lives as worse than some of those with lower incomes. A satiation point in life evaluations was also found in European countries at around €28K annually (Muresan et al., 2020).

This pattern of findings could partly depend on the choice of analytic strategy. In analyses of the same dataset as Jebb et al. (2018) but using lowess regression, researchers found no evidence of satiation or turning points in the relationship between income and people's life evaluations (Sacks et al., 2012; Stevenson and Wolfers, 2012). These conflicting results suggest that the effect of analytic strategy on results deserves a closer examination.

The Research Gap

While there has been much research on income and happiness, including according to how happiness is defined and measured, we are not away of any studies that have compared the relationship between income and happiness according to how income is defined and measured. We propose that the relationship between income and happiness may depend not only on how happiness is measured, but also on how income is measured and analyzed. To improve our knowledge of the relationship between income and happiness, this paper, we focus on nonlinearities in the relationship between income and happiness and re-analyze the ATUS data used by Kushlev et al. (2015) and Stone et al. (2018), as well as the GSOEP data used by Hudson et al. (2016). Specifically, while Kushlev et al. (2015) analyzed income as a continuous variable in the ATUS, we treat income the way it was measured: as a categorical variable. We compare these results to GSOEP data where we re-code the original continuous measure of income into categorical quantiles. To further explore nonlinearities in the relationship

between income and happiness, we also conduct local linear "lowess" and spline regression analyses.

We chose to re-analyze these data to address the question of differences in the relationship between income and happiness according to the measurement and analysis of income because the ATUS and GSOEP provide nationally representative data on people's feelings as experienced during specific "episodes" of the day after asking them to reconstruct what they did during the entire day. Thus, compared to data from Gallup, which measures affect "yesterday," measurements in the ATUS are more grounded in specific experiences, and therefore, less subject to recall bias (Kahneman et al., 2004). And unlike Gallup, which uses more crude, dichotomous ("yes-no") response scales, ATUS measures happiness along a standard seven-point Likert-type scale. In the GSOEP, we were also able to analyze data from the Experience Sampling Methodology (ESM), which asks people how they are feeling during specific episodes during the day and, as such, is even more grounded in specific experiences.

Measuring and Analyzing Income

The original ATUS income variable—family income—contains 16 uneven categories (see **Table 1**). For example, Category 11 has a range of ~$10K, whereas Category 14 has a range of ~$25K. The increasingly larger categories are designed to reflect declining marginal utility as an innate quality of income. Based on this, Kushlev et al. (2015) analyzed income as a continuous variable using the original uneven categories. Continuous scales, however, assume equal intervals between scale points—a strong assumption to make for the relatively arbitrary rate of change in the category ranges. Is increasing one's income from $20,000 to $25,000 really equidistant to increasing it from $35,000 to $40,000 (**Table 1**)? And can we really assume, for example, that adding $5,000 of additional income to $35,000 is the same as adding $10,000 of additional income to $40,000?

TABLE 1 | The original categories of income in the ATUS family income measure with number of individuals in each income category in the ATUS 2010, 2012, and 2013 well-being modules.

Group number	Income range	*N* (individuals)
1	Less than $5,000	883
2	$5,000–$7,499	645
3	$7,500–$9,999	903
4	$10,000–$12,499	1,221
5	$12,500–$14,999	1,096
6	$15,000–$19,999	1,773
7	$20,000–$24,999	2,005
8	$25,000–$29,999	1,989
9	$30,000–$34,999	2,044
10	$35,000–$39,999	1,809
11	$40,000–$49,999	2,959
12	$50,000–$59,999	2,831
13	$60,000–$74,999	3,466
14	$75,000–$99,999	4,011
15	$100,000–$149,999	3,706
16	$150,000 and over	2,635

Complete cases only for all variables analyzed.

Recognizing this issue, income researchers have adopted alternative strategies. For example, Stone et al. (2018) took the midpoints of each category of income, and then log-transformed it. Thus, they transformed the categorical measure of income into a continuous measure. This approach produced results for happiness consistent with the findings of Kushlev et al. (2015).

Both the increasing ranges of the income scale itself and its log-transformations reflect an assumed declining marginal utility of income: They treat a given amount of income increase at the higher end of the income distribution as having less utility than the same amount at the lower end of the distribution. But by subsuming income's declining utility in its very measurement (or transformation thereof), it becomes difficult to interpret a null relationship with happiness. In other words, we might not be seeing a declining marginal utility of income reflected on happiness because the income variable itself reflects its declining utility.

Even when the income variable itself does not reflect its declining utility, a null relationship between income and daily experiences of happiness has been observed. Hudson et al. (2016) used GSOEP, which contains a measure of income that is continuous in its original form. Whether analyzing this income measure in its raw original form or in transformed log and quadratic forms, a null relationship with happiness was observed. This approach, however, does not consider whether there might be nonlinear/log/quadratic turning or satiation points at higher levels of income—an issue also applicable to previous analyses of ATUS (Kushlev et al., 2015; Stone et al., 2018). This is important because there are theoretically both benefits and costs to achieving higher levels of income that could occur at various levels of income; however, this possibility has not yet been fully explored in ATUS or GSOEP data.

In sum, past research using ATUS has treated categorically measured income as a continuous variable, either assuming equidistance between scale points or attempting to create equidistance through statistical transformations. By doing so, however, researchers may have statistically accounted for the very utility of income for happiness that they are trying to test. In both ATUS and GSOEP, the question of whether there might be satiation and/or turning points at higher levels of income has not been fully considered. The present research explores whether treating income as a categorical variable in both ATUS and GSOEP would replicate past findings or reveal novel insights, focusing on possible nonlinearities in the relationship between income and happiness.

MATERIALS AND METHODS

Samples

We used data from ATUS well-being modules in 2010, 2012, and 2013. To facilitate future replications of this research, the ATUS extract builder was used to create the dataset (Hofferth et al., 2017).[1] The ATUS is a repeated cross-sectional survey

[1]https://www.atusdata.org

and is nationally representative of United States household residents aged 15 years and older. Its sampling frame is the Current Population Survey (CPS), which was conducted 2–5 months prior to the ATUS. Some items in the ATUS come from the CPS, including the household income item that we analyze.

Data from the GSOEP come from the Innovation Sample (IS), which is a subsample of the larger main GSOEP (Richter and Schupp, 2015). The main GSOEP and the IS are designed to be nationally representative. The IS contains information on household residents aged 17 years of age and older. We used two modules from these data: the 2012–2015 DRM module, which is a longitudinal survey, and the 2014–2015 ESM module.

Outcome Measures

In ATUS, participants were called on the phone and asked how they spent their time yesterday: what activities they were doing, for how long, who they spent time with and where they were located. This information was used to create their time use diary. A random selection of three activities were taken from these diaries and participants were asked how they felt during them. The feelings items were tired, sad, stressed, pain, and happy. Participants were also asked how meaningful what they were doing felt.

In GSOEP, participants were interviewed face to face for the DRM questions and through smartphones for the ESM questions. In the DRM, as in the ATUS, they were asked how they spent their time yesterday and, for a random selection of three activities, they were asked further details about how they felt. In the ESM, participants were randomly notified on mobile phones at seven random points during the day for around 1 week. As in the DRM, they were asked how they were spending their time at the point of notification, as well as how they felt. Participants in both ESM and DRM samples were asked about whether they were feeling happy, as well as other emotions such as sadness, stress, and boredom.

The focus of this research is on the happiness items from both the ATUS and GSOEP to highlight differences according to the treatment of the independent measure of income rather than differences according to the dependent outcome of emotional well-being.

Analyses

Data were analyzed in STATA 15 and jamovi. The **Supplementary Material S1** file contains the STATA command file for the main commands written to analyze the data. In both ATUS and GSOEP, OLS regressions were conducted with happiness as the outcome measure and income as the explanatory measure. Following Kushlev et al. (2015) and Hudson et al. (2016), the average happiness across all activities each day was taken to create an individual-level measure. Because the GSOEP DRM sample contained multiple observations across years, the SEs were clustered at the individual level for models using this dataset.

The treatment of income differed according to the dataset because income was collected differently in each dataset. In the ATUS, income was first analyzed in continuous, log, and quadratic forms in OLS regressions, as in other research (Kushlev et al., 2015; Hudson et al., 2016). Next, it was analyzed as a categorical variable with 16 categories, preserving the identical format that it was originally collected in from the CPS questionnaire.

TABLE 2 | The range and number of person-year observations of the GSOEP Income 4 variable divided into 16 quantiles.

Quantile number	Income minimum	Income maximum	*N* (observations)
1	2,400	11,520	433
2	11,616	14,400	459
3	14,472	18,000	584
4	18,024	19,200	228
5	19,356	21,600	427
6	21,840	24,000	520
7	24,120	26,880	306
8	26,940	30,000	660
9	30,240	32,400	257
10	33,000	36,000	631
11	36,360	38,400	193
12	39,000	42,000	430
13	42,600	48,000	539
14	49,032	54,000	289
15	54,720	64,800	400
16	66,000	360,000	410

Complete cases only for all variables analyzed.

In GSOEP, the income variable in the dataset is provided in continuous form because participants reported their monthly income as an integer. To compare to the ATUS results, 16 quantiles of income were created and analyzed in GSOEP DRMs (see **Table 2** - note that there were insufficient observations to conduct these analyses with GSOEP ESMs). This income variable was also analyzed in continuous, log, and quadratic forms.

Omnibus *F*-tests and effect sizes (n^2) are also reported to compare the categorical, continuous, log, and quadratic approaches.

We conducted lowess and spline regressions to further investigate possible nonlinearities in the relationship between income and happiness. For the lowess regressions, the smoothing parameter was set at of 0.08. For the regression splines, we fitted knots at four quartiles and five quantiles of income. We also used the results of OLS regressions treating income as a categorical variable, as well as the results of the lowess regression treating income as continuous, to fit knots at pre-specified values of income (where these analyses suggested there could be turning and/or satiation points).

Complete case analyses were conducted with 33,976 individuals in ATUS, 6,766 individuals in German DRMs, and 249 individuals in German ESMs. There was item-missing data in some samples (ATUS, 1.7% missing; GSOEP DRMs, 8.2% missing; GSOEP ESMs data, and 6.0% missing). We make analytical and not population inferences and therefore do not use survey weights (Pfeffermann, 1996).

Controls

Results are presented without and with controls for demographic and diary characteristics. Following Kushlev et al. (2015), Hudson et al. (2016), and Stone et al. (2018), these controls were age, gender, marital status, ethnic

TABLE 3 | List of variables used in analyses in ATUS and GSOEP.

Variable	ATUS	GSOEP
Happiness	x	x
Income		
Continuous	x	x
Log	x	x
Quadratic	x	x
Categorical	x	x
Age	x	x
Gender	x	x
Marital status	x	x
Ethnic background		
Hispanic/Black	x	
German origin		x
Health		
Physical or cognitive difficulty	x	
Self-rated general health		x
Employment status	x	x
Children		
Children <18 years in household	x	
Number of children		x
Diary day was weekend	x	x
Year of survey	x	

background,[2] health,[3] employment status, children,[4] and whether the day was a weekend. We also control for the year of the survey in ATUS DRM data to address the issue that our results are not due to new data but rather how we treat the income variable.

The list of variables we use in analyses are in **Table 3**.

RESULTS

In both ATUS and GSOEP, daily happiness was analyzed using a 0–6 scale (in GSOEP scale points 1–7 were recoded to 0–6 to match ATUS). The ATUS mean happiness was 4.38 (SD = 1.33). The GSOEP DRM mean happiness was 2.91 (SD = 1.46), and the GSOEP ESM mean happiness was 2.65 (SD = 1.03).

Magnitude

The magnitude of our results can be considered in the context of effect sizes from other research on demographic characteristics and daily happiness (Kahneman et al., 2004; Stone et al., 2010; Luhmann et al., 2012; Hudson et al., 2019). For example, the effect size for the relationship between age and daily experiences of happiness was 0.16 in Stone et al. (2010). Our effect sizes range from 0.06 to 0.37. Throughout, we focus on coefficients, their 95% CIs, and visualizations of these coefficients and CIs, rather than on their statistical significance (Lakens, 2021). The purpose of this is to highlight how analytic treatments of income affect the magnitude and precision of the relationship between income and happiness.

ATUS-DRM

When treating the 16-category family income variable as continuous in OLS regressions, there was no substantive relationship between income and happiness as in other prior research (Kushlev et al., 2015; Hudson et al., 2016; Stone et al., 2018). Out of the linear, squared, and log coefficients without and with controls, the largest and most precise coefficients were with controls; for linear income it was ($b = -0.006$, 95% CI = −0.01, −0.002), squared income ($b = -0.0001$, 95% CI = 0.0003, 0.00006), and log income ($b = -0.03$, 95% CI = −0.05, 0.001). The omnibus F-test (without controls) for linear income was $F = 0.28$, $n^2 = 0.000008$ (95% CI = 0.00, 0.0002), for income squared was $F = 1.60$, $n^2 = 0.00005$ (95% CI = 0.00, 0.0003), and for log income was $F = 0.23$, $n^2 = 0.000006$ (95% CI = 0.00,0.0002).

The categorization of income focused attention on those with incomes of $35–40K, who appeared substantively happier than some of those with higher incomes (and lower incomes; see **Figure 1**). For example, with controls, those with incomes of $35–40K appeared happier relative to those with incomes of $150K+ ($b = 0.16$, 95% CI: 0.08, 0.24) and $100–150K ($b = 0.14$, 95% CI: 0.07, 0.221). The omnibus test for categorical income was $F = 1.61$, $n^2 = 0.007$ (95% CI = 0.00, 0.0009).

Results from regression splines and a lowess regression suggested null results overall (see **Figure 2**). Further details of the analyses are in **Supplementary Material S2**.

GSOEP-DRM

When treating the continuous household income variable as continuous (in €10,000s) in OLS regressions, there was no substantive relationship between income and happiness as in

[2]In the ATUS this was Hispanic and Black, in GSOEP this was German origin.
[3]In the ATUS this was whether the respondent had any physical or cognitive difficulty (yes/no), in GSOEP this was self-rated general health (bad, poor, satisfactory, good, and very good).
[4]In the ATUS this was presence of children <18 years in the household, in GSOEP this was number of children.

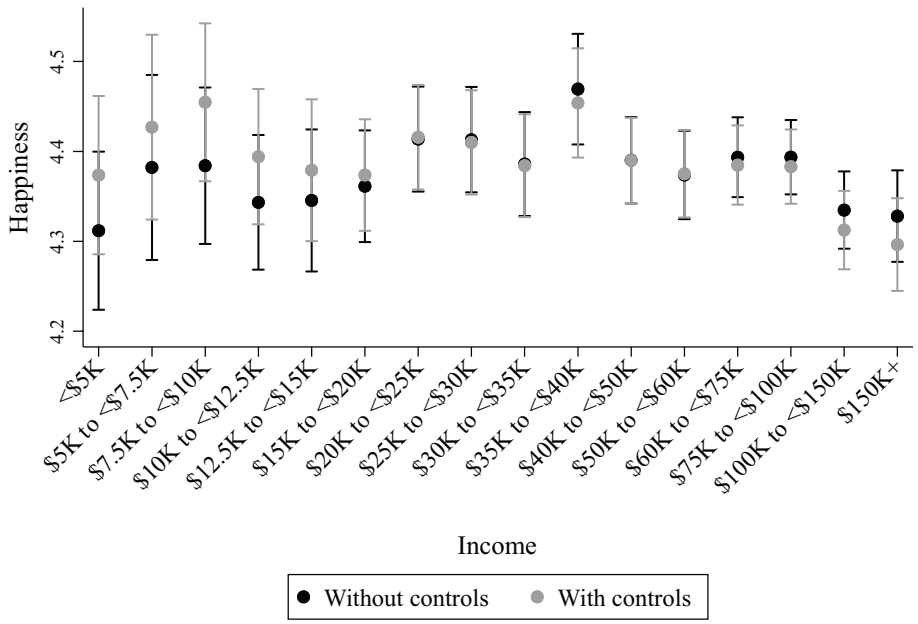

FIGURE 1 | Predicted values of average individual happiness in the American Time Use Survey (ATUS) at the 16 values of the family income variable without and with controls. Covariates at means. 95% CI.

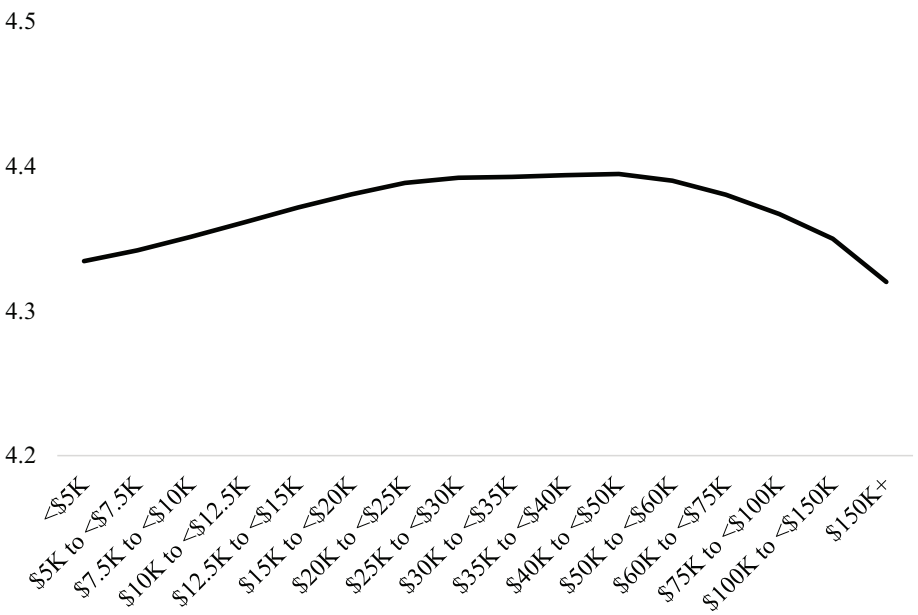

FIGURE 2 | Line graph of predicted values from lowess regressions explaining variance in happiness from income treated as a continuous variable in ATUS.

other prior research (Kushlev et al., 2015; Hudson et al., 2016; Stone et al., 2018). The association with the largest magnitude and most precision was for log income with controls ($b = -0.08$, 95% CI = −0.18, 0.01).[5]

[5]This association was stronger and more precise when equivalizing income (dividing by the square root of household size), $b = -0.16$, 95%CI = −0.06, −0.27, underscoring the importance of transparency in the treatment of income.

As in ATUS, treating the variable as categorical suggested some relationships between income and happiness. These results drew attention to those third quantile (~€14–18K), who seemed happier than those both higher and lower in income (see **Figure 3**). For example, with controls, they were happier than those in quantiles 13 (€42.6–48K, $b = 0.46$, 95% CI = 0.25, 0.67), seven (~€24–27K, $b = 0.34$, 95% CI = 0.13, 0.56), and one (€2.40–11,520K, $b = 0.28$, 95% CI = 0.05, 0.51).

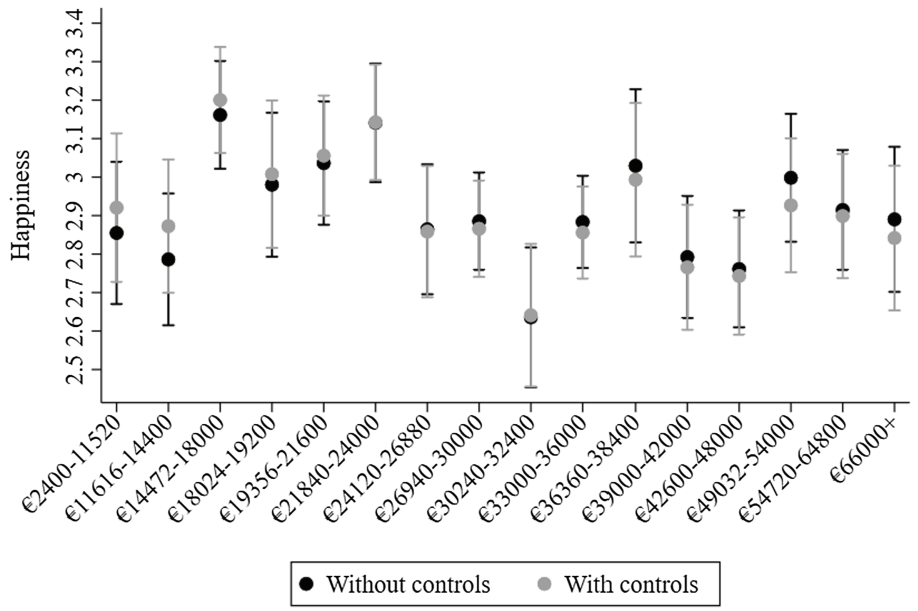

FIGURE 3 | Predicted values of average person-year happiness from GSOEP DRMs at 16 quantiles of income (Income 4) without and with controls. Covariates at means. 95% CI.

The omnibus test for categorical income was $F = 4.00$, $n^2 = 0.009$ (95% CI = 0.003, 0.01), whereas the omnibus test for linear income was $F = 0.09$, $n^2 = 0.00001$ (95% CI = 0.00, 0.0007). The omnibus for log income was $F = 1.42$, $n^2 = 0.0002$ (95% CI = 0.00, 0.0001) and for income squared it was $F = 0.96$, $n^2 = 0.0001$ (95% CI = 0.00, 0.001).

The lowess and spline regressions suggested null results overall, as the coefficients were small in magnitude (see **Figure 4**). Further details of the analyses are in **Supplementary Material S3**.

GSOEP-ESM

There was no evidence to suggest any substantive association between income and happiness in ESM data for linear income, income squared, log income, in the lowess regressions, or regression splines. A visualization of the lowess results are in **Figure 5** and further details of the analyses are in **Supplementary Material S4**.

The omnibus *F*-test for linear income was $F = 0.53$, $n^2 = 0.002$ (95%CI = −0.00, 0.03), and for log income it was $F = 0.12$, $n^2 = 0.0005$, 95%CI = 0.00, 0.02. For income squared it was $F = 0.63$, $n^2 = 0.003$, 95%CI = 0.00 0.03.

DISCUSSION

Is income creating a signal in these data on daily experiences of happiness, or is it all simply noise? The present results suggest that whether income can be concluded as being associated with daily experiences of happiness may depend on how income is analyzed. When income in ATUS is analyzed in its original, categorical form, there is some evidence that some people with higher incomes feel somewhat less happy than some of those with lower incomes. When the continuous income variable in GSOEP is split into categories, a similar pattern is observed. This is not inconsistent with the findings of Kushlev et al. (2015), Hudson et al. (2016), and Stone et al. (2018), who found no relationship between income and daily feelings of happiness in the same data when income was analyzed as a continuous variable. It simply illustrates that a relationship between income and happiness could be interpreted when treating income categorically rather than continuously.

There are at least three possible interpretations to our overall results. One interpretation tends toward conservative. We conducted multiple comparisons of many transformations of income, which might inspire some to question whether we should have accounted for this in some way by adjusting for multiple comparisons. Although we found some evidence of differences in happiness according to income, such an adjustment might lead to an overall null conclusion when characterizing the relationship between income on happiness. A second interpretation is more generous. Within this perspective, one might emphasize the fact that because our income measures were correlated, no correction for multiple comparisons was required. It could then be argued that because we found some evidence for the relationship between income on happiness, there is good evidence that the overall effect is not null. A more moderate perspective, and the one adopted in this paper, is that because the overall pattern of our results showed mixed null and nonnull results, we can make an overall conclusion of some differences in happiness according to income. We also noticed that equivalizing income in the German data strengthened the relationship of income and happiness, further supporting the conclusion

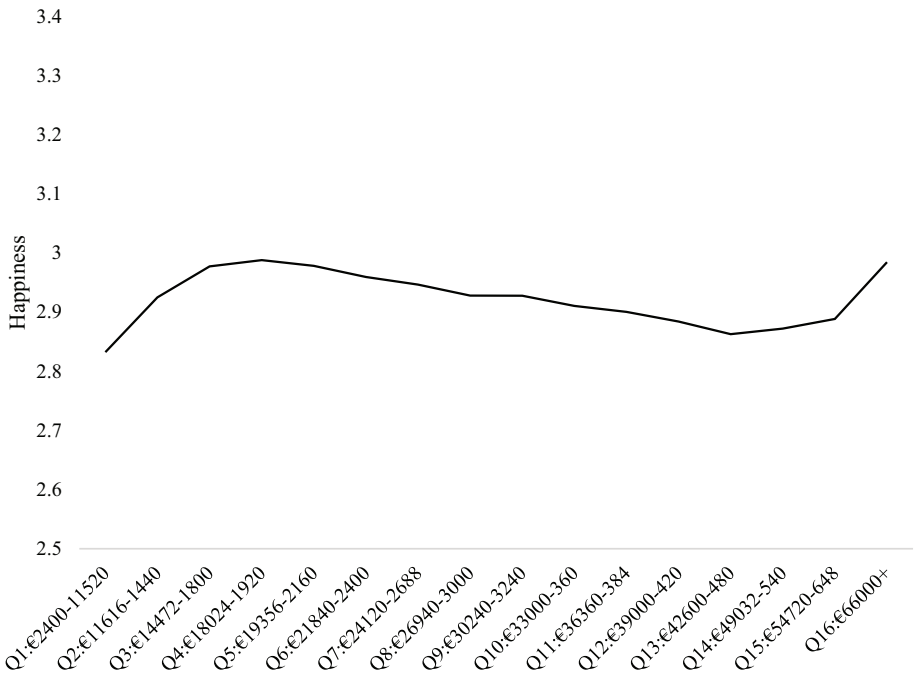

FIGURE 4 | Line graph of predicted values from lowess regressions explaining variance in happiness from income treated as a continuous variable in GSOEP DRMs at 16 quantiles of income.

of some differences—and that the analytic treatment of income matters.

Based on the moderate perspective, we conclude that there is very little evidence of any relationship between income and daily experiences of happiness—and any relationship that does exist would suggest higher income could be associated with less happiness. The results do not support the results of Sacks et al. (2012) or Killingsworth (2021), where a greater income was associated with greater happiness, and there were no satiation or turning points (see also Stevenson and Wolfers, 2012). These results are more aligned with Kahneman and Deaton (2010), who found a satiation point in the relationship between income daily experiences of happiness, researchers finding no association between income and happiness (Kushlev et al., 2015; Jebb et al., 2018; Casinillo et al., 2020, 2021), who found that higher income can be associated with worse evaluations of life. We suggest the analytic strategy for income could contribute to explaining discrepant results in existing literature, and researchers should be clear about the approaches they have tested, although we acknowledge that sampling differences could play a role, too.

Overall, the results were broadly consistent between countries because there was no substantive relationship between income and happiness when income was treated continuously but there appeared to be relationships when treating income categorically. Despite a similar overall pattern in the income results, there were other difference between countries. German residents rated their happiness as lower than United States residents (a difference of ~1.5 scale points out of seven). This could be because of different interpretations of the word "happiness" in Germany and the United States. The word for happiness in German used in the survey—*glück*—can mean something more akin to lucky or optimistic—which is different from the meaning of word "happy" in the United States. Despite this linguistic difference, those with higher incomes were still less happy than some of those with lower incomes in both samples.

Limitations

One limitation to our results is the representativeness of the income distribution. Household surveys like those that we used do not tend to capture the "tails" of the income distribution very well: People in institutions and without addresses are excluded from these sample populations, which omits populations such as those living in nursing homes and prisons, as well as the homeless. Moreover, people do not always self-report their income accurately due to issues such as social desirability bias (Angel et al., 2019). Existing studies that have focused on those with very low incomes do tend to find that low income is associated with low happiness (Diener and Biswas-Diener, 2002; Clark et al., 2016; Adesanya et al., 2017). In ATUS, the highest household income value available was $150K, whereas in GSOEP it was €360K. Thus, it is not always clear whether the very affluent, such as millionaires, are represented in these samples (Smeets et al., 2020). Overall, our results cannot be taken as representative of people who are very poor or rich and should not be interpreted as such.

Another limitation is that the present results cannot be interpreted casually because there has been no manipulation of income in these data nor exploration of mechanisms and there was no longitudinal data in ATUS. As discussed by

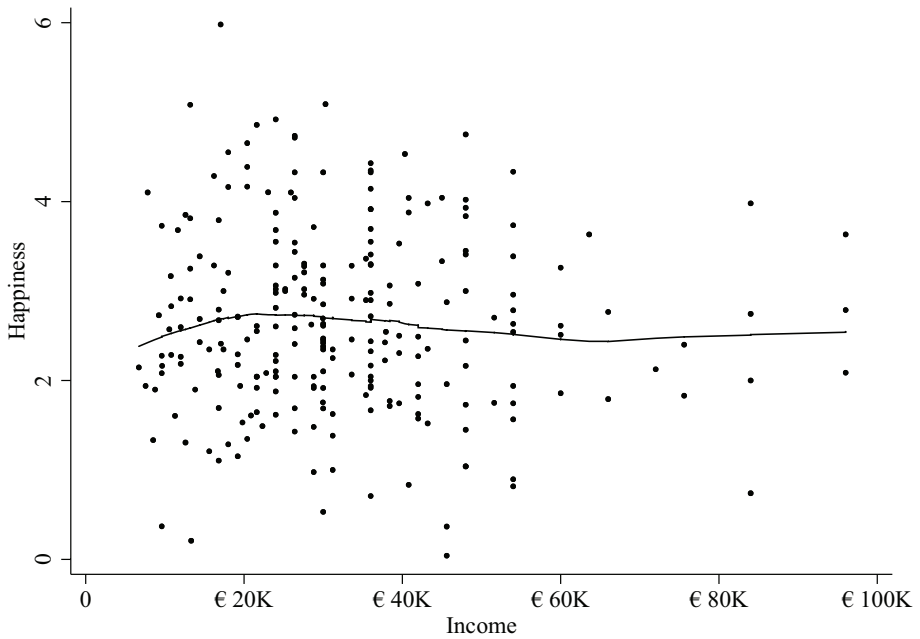

FIGURE 5 | Results of local linear "lowess" regression from GSOEP Experience Sampling Methodology (ESM) data with happiness as the outcome and continuous annual income as the explanatory variable.

Kushlev et al. (2015), there are issues such as reverse causality. Here, however, some of our results potentially suggest an alternative reverse causality pathway, whereby less happy people may select into earning more income. Because the counterfactual is not apparent—we do not know how happy people with high incomes would be without their higher income—it could also be that those with high incomes would be even less happy than they currently are if they had not attained their current level of income. In other words, people with high incomes may have started out as less happy in the first place and be even less happy if they did not have high incomes.

A further limitation is the time period of the data, especially that they were collected prior to the COVID-19 pandemic. This could be an issue because it is possible that the relationship between income and daily experiences of happiness has changed, such as due to the exacerbation of health inequalities and restrictions on freedom of movement due to nationwide lockdowns. Our study does not provide any information on the longer-term and health and well-being consequences of both COVID-19 itself and the policy response to COVID-19 (Aknin et al., 2022). As one example, access to green space, which has health and well-being benefits, is lower among those with low income, and this mechanism between income and happiness may have become more salient during COVID-19 (Geary et al., 2021). Overall, it is important to consider the regional, political, and socioeconomic contexts in which income is attained to understand its relationship with well-being, including levels of income in reference groups such as neighbors, friends, and colleagues (Luttmer, 2005; De Neve and Sachs, 2020). It would be important to replicate the results in this research with more recent data to address the limitation that the data we used are not recent, considering our broader point that the measurement and analysis of income should be considered as carefully as the measurement and analysis of happiness.

Future Directions

This research points to several directions for future research. One direction relates to data and measures: Nonlinearities in the relationship between income and happiness could be examined using time use data from other countries, considered between countries and/or within countries over time (Deaton et al., 2008; De Neve et al., 2018), and investigated for measures of emotional states other than happiness (Piff and Moskowitz, 2018). In general, our results suggest that researchers should pay attention to how income is measured and analyzed when considering how it is related to happiness, which complements findings from other research that the way happiness is measured and analyzed is important (Kahneman and Deaton, 2010; Jebb et al., 2018).

Future research could also explore mechanisms that may explain our findings. In addition to those mentioned in the Introduction—expectations (Graham and Pettinato, 2002; Nickerson et al., 2003), time use (Aguiar and Hurst, 2007; Hamermesh and Lee, 2007; Bianchi and Vohs, 2016; Nikolaev, 2018; Sharif et al., 2021); generosity (Dunn et al., 2008; Kraus et al., 2010; Piff et al., 2010; Aknin et al., 2012; Balakrishnan et al., 2017; Macchia and Whillans, 2022), and sense of self (Snibbe and Markus, 2005; Stephens et al., 2007)—another is the identity-related effect of transitioning between socioeconomic groups. Though one might expect upward mobility to

be associated with greater happiness, research suggests that some working class people do not wish to become upwardly mobile because it could lead to a loss of identity and change in community (Akerlof, 1997; Friedman, 2014). Indeed, upward intergenerational mobility is associated with worse life evaluations in the United Kingdom—though not in Switzerland (Hadjar and Samuel, 2015), although recent findings show substantial negative effects of downward mobility, too (Dolan and Lordan, 2021). Over time, therefore, the degree of mobility in a population could influence the relationship between income and happiness in both positive and negative directions.

Additionally, social comparisons could drive the effects of higher income on happiness. Higher income might not benefit happiness if one's reference group—that is, the people to whom we compare or have knowledge of in some form (Hyman, 1942; Shibutani, 1955; Runciman, 1966)—changes with higher socioeconomic status. As income increases, people might compare themselves to others who are also doing similarly or better to them, and then not feel or think that they are doing any better by comparison—or even feel worse (Cheung and Lucas, 2016). This is one of the explanations for the well-known "Easterlin Paradox" (Easterlin, 1974), which suggests that as national income rises people do not become happier because they compare their achievements to others. The paradox is debated (Sacks et al., 2012). Additionally, some research shows that it is possible to view others' greater success as one's own future opportunity and for upward social comparisons to then positively impact upon well-being (Senik, 2004; Davis and Wu, 2014; Ifcher et al., 2018). As with the role of mobility in the relationship between income and happiness, it is unclear whether the role of social comparisons would create a positive or negative impact over time and future research could explore this.

Final Remarks

Overall, our results provide some evidence that individual attainment in terms of income may not equate to the attainment of individual happiness—and could even be associated with less daily happiness, depending upon how income is measured and analyzed. These results suggest that how income is associated with happiness depends on how income is measured and analyzed. They provide some support to the idea that financial achievement can have both costs and benefits, potentially informing normative discussions about the optimal distribution of income in society.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found at: https://www.atusdata.org (The ATUS extract builder was used to create the ATUS dataset, see Hofferth et al., 2017). GSOEP data were requested from https://www.diw.de/en/diw_02.c.222516.en/data.html, see Richter and Schupp, 2015.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

LK and KK contributed to conception and design of the study. LK organized the data, performed the statistical analysis in STATA, and wrote the first draft of the manuscript. KK performed additional statistical analysis in jamovi and wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

LK was supported by a London School of Economics PhD scholarship during early work and later by the National Institute for Health Research (NIHR) Applied Research Collaboration (ARC) West Midlands. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

ACKNOWLEDGMENTS

LK thanks Professor Paul Dolan and Dr Georgios Kavetsos for their support early on in conducting this research, as well as Professor Richard Lilford for insights about multiple comparisons.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.883137/full#supplementary-material

REFERENCES

Adesanya, O., Rojas, B. M., Darboe, A., and Beogo, I. (2017). Socioeconomic differential in self-assessment of health and happiness in 5 African countries: finding from world value survey. *PLoS One* 12:e0188281. doi: 10.1371/journal.pone.0188281

Aguiar, M., and Hurst, E. (2007). Measuring trends in leisure: the allocation of time over five decades. *Q. J. Econ.* 122, 969–1006. doi: 10.1162/qjec.122.3.969

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica* 65, 1005–1027. doi: 10.2307/2171877

Aknin, L. B., De Neve, J. E., Dunn, E. W., Fancourt, D. E., Goldberg, E., Helliwell, J. F., et al. (2022). Mental health during the first year of the COVID-19 pandemic: a review and recommendations for moving forward. *Perspect. Psychol. Sci.* 19:17456916211029964. doi: 10.1177/17456916211029964

Aknin, L. B., Hamlin, J. K., and Dunn, E. W. (2012). Giving leads to happiness in young children. *PLoS One* 7:e39211. doi: 10.1371/journal.pone.0039211

Aknin, L. B., Norton, M. I., and Dunn, E. W. (2009). From wealth to well-being? Money matters, but less than people think. *J. Posit. Psychol.* 4, 523–527. doi: 10.1080/17439760903271421

Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *Econ. J.* 100, 464–477. doi: 10.2307/2234133

Angel, S., Disslbacher, F., Humer, S., and Schnetzer, M. (2019). What did you really earn last year? Explaining measurement error in survey income data. *J. R. Stat. Soc. A. Stat. Soc.* 182, 1411–1437. doi: 10.1111/rssa.12463

Angner, E. (2010). Subjective well-being. *J. Socio-Econ.* 39, 361–368. doi: 10.1016/j.socec.2009.12.001

Balakrishnan, A., Palma, P. A., Patenaude, J., and Campbell, L. (2017). A 4-study replication of the moderating effects of greed on socioeconomic status and unethical behaviour. *Sci. Data* 4:160120. doi: 10.1038/sdata.2016.120

Becchetti, L., Pelloni, A., and Rossetti, F. (2008). Relational goods, sociability, and happiness. *Kyklos* 61, 343–363. doi: 10.1111/j.1467-6435.2008.00405.x

Bianchi, E. C., and Vohs, K. D. (2016). Social class and social worlds: income predicts the frequency and nature of social contact. *Soc. Psychol. Personal. Sci.* 7, 479–486. doi: 10.1177/1948550616641472

Casinillo, L. F., Casinillo, E. L., and Aure, M. R. K. L. (2021). Economics of happiness: a social study on determinants of well-being among employees in a state university. *Philippine Soc. Sci. J.* 4, 42–52. doi: 10.52006/main.v4i1.316

Casinillo, L. F., Casinillo, E. L., and Casinillo, M. F. (2020). On happiness in teaching: an ordered logit modeling approach. *JPI* 9, 290–300. doi: 10.23887/jpi-undiksha.v9i2.25630

Cheung, F., and Lucas, R. E. (2016). Income inequality is associated with stronger social comparison effects: the effect of relative income on life satisfaction. *J. Pers. Soc. Psychol.* 110, 332–341. doi: 10.1037/pspp0000059

Clark, A. E., D'Ambrosio, C., and Ghislandi, S. (2016). Adaptation to poverty in long-run panel data. *Rev. Econ. Stat.* 98, 591–600. doi: 10.1162/REST_a_00544

Davis, L., and Wu, S. (2014). Social comparisons and life satisfaction across racial and ethnic groups: the effects of status, information and solidarity. *Soc. Indic. Res.* 117, 849–869. doi: 10.1007/s11205-013-0367-y

De Neve, J. E., and Sachs, J. D. (2020). The SDGs and human well-being: a global analysis of synergies, trade-offs, and regional differences. *Sci. Rep.* 10, 1–12. doi: 10.1038/s41598-020-71916-9

De Neve, J. E., Ward, G., De Keulenaer, F., Van Landeghem, B., Kavetsos, G., and Norton, M. I. (2018). The asymmetric experience of positive and negative economic growth: global evidence using subjective well-being data. *Rev. Econ. Stat.* 100, 362–375. doi: 10.1162/REST_a_00697

Deaton, A. (2008). Income, health, and well-being around the world: evidence from the Gallup world poll. *J. Econ. Perspect.* 22, 53–72. doi: 10.1257/jep.22.2.53

Di Tella, R., Haisken-De New, J., and MacCulloch, R. (2010). Happiness adaptation to income and to status in an individual panel. *J. Econ. Behav. Organ.* 76, 834–852. doi: 10.1016/j.jebo.2010.09.016

Diener, E., and Biswas-Diener, R. (2002). Will money increase subjective well-being? *Soc. Indic. Res.* 57, 119–169. doi: 10.1023/A:1014411319119

Dolan, P., and Kudrna, L. (2016). "Sentimental hedonism: pleasure, purpose, and public policy" in *International Handbooks of Quality-of-Life. Handbook of Eudemonic Well-Being*. ed. J. Vittersø (Springer International Publishing AG), 437–452.

Dolan, P., and Lordan, G. (2021). Climbing up ladders and sliding down snakes: an empirical assessment of the effect of social mobility on subjective wellbeing. *Rev. Econ. Househ.* 19, 1023–1045. doi: 10.1007/s11150-020-09487-x

Dolan, P., Peasgood, T., and White, M. (2008). Do we realy know what makes us happy? A review of the economic literaure on the factors associated with subjective well-being. *J. Econ. Psychol.* 29, 94–122. doi: 10.1016/j.joep.2007.09.001

Dunn, E. W., Aknin, L. B., and Norton, M. I. (2008). Spending money on others promotes happiness. *Science* 319, 1687–1688. doi: 10.1126/science.1150952

Easterlin, R. A. (1974). "Does economic growth improve the human lot? Some empirical evidence," in *Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz*. eds. P. A. David and M. W. Reder (New York: Academic Press, Inc.).

Friedman, S. (2014). The price of the ticket: rethinking the experience of social mobility. *Sociology* 48, 352–368. doi: 10.1177/0038038513490355

Geary, R. S., Wheeler, B., Lovell, R., Jepson, R., Hunter, R., and Rodgers, S. (2021). A call to action: improving urban green spaces to reduce health inequalities exacerbated by COVID-19. *Prev. Med.* 145:106425. doi: 10.1016/j.ypmed.2021.106425

Graham, C., and Pettinato, S. (2002). Frustrated achievers: winners, losers and subjective well-being in new market economies. *J. Dev. Stud.* 38, 100–140. doi: 10.1080/00220380412331322431

Hadjar, A., and Samuel, R. (2015). Does upward social mobility increase life satisfaction? A longitudinal analysis using British and Swiss panel data. *Res. Soc. Stratif. Mobil.* 39, 48–58. doi: 10.1016/j.rssm.2014.12.002

Hamermesh, D. S., and Lee, J. (2007). Stressed out on four continents: time crunch or yuppie kvetch? *Rev. Econ. Stat.* 89, 374–383. doi: 10.1162/rest.89.2.374

Harsanyi, J. C. (1997). Utilities, preferences, and substantive goods. *Soc. Choice Welf.* 14, 129–145.

Helliwell, J. F., and Putnam, R. D. (2004). The social context of well–being. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 359, 1435–1446. doi: 10.1098/rstb.2004.1522

Hofferth, S., Flood, S., and Sobek, M. (2017). American time use survey data extract system: version 26 [machine-readable database]. College Park, MD: University of Maryland and Minneapolis, MN: University of Minnesota.

Hudson, N. W., Lucas, R. E., and Donnellan, M. B. (2019). Healthier and happier? A 3-year longitudinal investigation of the prospective associations and concurrent changes in health and experiential well-being. *Personal. Soc. Psychol. Bull.* 45, 1635–1650. doi: 10.1177/0146167219838547

Hudson, N. W., Lucas, R. E., Donnellan, M. B., and Kushlev, K. (2016). Income reliably predicts daily sadness, but not happiness: a replication and extension of Kushlev, Dunn, and Lucas (2015). *Soc. Psychol. Personal. Sci.* 7, 828–836. doi: 10.1177/1948550616657599

Hyman, H. H. (1942). "The psychology of status," in *Archives of Psychology* (Columbia University).

Ifcher, J., Zarghamee, H., and Graham, C. (2018). Local neighbors as positives, regional neighbors as negatives: competing channels in the relationship between others' income, health, and happiness. *J. Health Econ.* 57, 263–276. doi: 10.1016/j.jhealeco.2017.08.003

Jebb, A. T., Tay, L., Diener, E., and Oishi, S. (2018). Happiness, income satiation and turning points around the world. *Nat. Hum. Behav.* 2, 33–38. doi: 10.1038/s41562-017-0277-0

Kahneman, D., and Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. *Proc. Natl. Acad. Sci. U. S. A.* 107, 16489–16493. doi: 10.1073/pnas.1011492107

Kahneman, D., Krueger, A., Schkade, D., Schwarz, N., and Stone, A. (2004). A survey method for characterizing daily life experience: the day reconstruction method. *Science* 306, 1776–1780. doi: 10.1126/science.1103572

Killingsworth, M. A. (2021). Experienced well-being rises with income, even above $75,000 per year. *Proc. Natl. Acad. Sci. U. S. A.* 118:e2016976118. doi: 10.1073/pnas.2016976118

Kraus, M. W., Côté, S., and Keltner, D. (2010). Social class, contextualism, and empathic accuracy. *Psychol. Sci.* 21, 1716–1723. doi: 10.1177/0956797610387613

Kushlev, K., Dunn, E. W., and Lucas, R. E. (2015). Higher income is associated with less daily sadness but not more daily happiness. *Soc. Psychol. Personal. Sci.* 6, 483–489. doi: 10.1177/1948550614568161

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspect. Psychol. Sci.* 16, 639–648. doi: 10.1177/1745691620958012

Layard, R., Mayraz, G., and Nickell, S. (2008). The marginal utility of income. *J. Public Econ.* 92, 1846–1857. doi: 10.1016/j.jpubeco.2008.01.007

Luhmann, M., Hofmann, W., Eid, M., and Lucas, R. E. (2012). Subjective well-being and adaptation to life events: a meta-analysis. *J. Pers. Soc. Psychol.* 102, 592–615. doi: 10.1037/a0025948

Luttmer, E. F. (2005). Neighbors as negatives: relative earnings and well-being. *Q. J. Econ.* 120, 963–1002. doi: 10.1162/003355305774268255

Macchia, L., and Whillans, A. V. (2022). The link between income, income inequality, and prosocial behavior around the world. *Soc. Psychol.* 52, 375–386. doi: 10.1027/1864-9335/a000466

Muresan, G. M., Ciumas, C., and Achim, M. V. (2020). Can money buy happiness? Evidence for European countries. *Appl. Res. Qual. Life* 15, 953–970. doi: 10.1007/s11482-019-09714-3

Nickerson, C., Schwarz, N., Diener, E., and Kahneman, D. (2003). Zeroing in on the dark side of the American dream: a closer look at the negative consequences of the goal for financial success. *Psychol. Sci.* 14, 531–536. doi: 10.1046/j.0956-7976.2003.psci_1461.x

Nikolaev, B. (2018). Does higher education increase hedonic and eudaimonic happiness? *J. Happiness Stud.* 19, 483–504. doi: 10.1007/s10902-016-9833-y

Nussbaum, M. C. (2008). Who is the happy warrior? Philosophy poses questions to psychology. *J. Leg. Stud.* 37, S81–S113. doi: 10.1086/587438

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Stat. Methods Med. Res.* 5, 239–261. doi: 10.1177/096228029600500303

Piff, P. K., Kraus, M. W., Côté, S., Cheng, B. H., and Keltner, D. (2010). Having less, giving more: the influence of social class on prosocial behavior. *J. Pers. Soc. Psychol.* 99, 771–784. doi: 10.1037/a0020092

Piff, P. K., and Moskowitz, J. P. (2018). Wealth, poverty, and happiness: social class is differentially associated with positive emotions. *Emotion* 18, 902–905. doi: 10.1037/emo0000387

Richter, D., and Schupp, J. (2015). The SOEP innovation sample (SOEP IS). *Schmollers Jahr.* 135, 389–399. doi: 10.3790/schm1353389

Runciman, W. (1966). *Relative Deprivation, Social Justice: Study Attitudes Social Inequality in 20th Century England.* Berkeley: University of California Press.

Sacks, D. W., Stevenson, B., and Wolfers, J. (2012). The new stylized facts about income and subjective well-being. *Emotion* 12, 1181–1187. doi: 10.1037/a0029873

Sen, A. (1999). *Development as Freedom*. New York: Alfred A. Knopf.

Senik, C. (2004). When information dominates comparison: learning from Russian subjective panel data. *J. Public Econ.* 88, 2099–2123. doi: 10.1016/S0047-2727(03)00066-5

Sharif, M. A., Mogilner, C., and Hershfield, H. E. (2021). Having too little or too much time is linked to lower subjective well-being. *J. Pers. Soc. Psychol.* 121, 933–947.

Shibutani, T. (1955). Reference groups as perspectives. *Am. J. Sociol.* 60, 562–569. doi: 10.1086/221630

Smeets, P., Whillans, A., Bekkers, R., and Norton, M. I. (2020). Time use and happiness of millionaires: evidence from the Netherlands. *Soc. Psychol. Personal. Sci.* 11, 295–307. doi: 10.1177/1948550619854751

Snibbe, A. C., and Markus, H. R. (2005). You can't always get what you want: educational attainment, agency, and choice. *J. Pers. Soc. Psychol.* 88, 703–720. doi: 10.1037/0022-3514.88.4.703

Stephens, N. M., Markus, H. R., and Townsend, S. (2007). Choice as an act of meaning: the case of social class. *J. Pers. Soc. Psychol.* 93, 814–830. doi: 10.1037/0022-3514.93.5.814

Stevenson, B., and Wolfers, J. (2012). Subjective well-being and income: is there any evidence of satiation? *Am. Econ. Rev.* 103, 598–604. doi: 10.1257/aer.103.3.598

Stone, A., Schneider, S., Krueger, A., Schwartz, J. E., and Deaton, A. (2018). Experiential wellbeing data from the American time use survey: comparisons with other methods and analytic illustrations with age and income. *Soc. Indic. Res.* 136, 359–378. doi: 10.1007/s11205-016-1532-x

Stone, A. A., Schwartz, J. E., Broderick, J. E., and Deaton, A. (2010). A snapshot of the age distribution of psychological well-being in the United States. *Proc. Natl. Acad. Sci. U. S. A.* 107, 9985–9990. doi: 10.1073/pnas.1003744107

Sunstein, C. R. (2021). Some costs and benefits of cost-benefit analysis. *Daedalus* 150, 208–219. doi: 10.1162/daed_a_01868

Tiberius, V. (2006). Well-being: psychological research for philosophers. *Philos. Compass* 1, 493–505. doi: 10.1111/j.1747-9991.2006.00038.x

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 | Frontiers in Psychology

SYSTEMATIC REVIEW
published: 01 June 2022
doi: 10.3389/fpsyg.2022.840187

Evaluation of Psychometric Properties of Hardiness Scales: A Systematic Review

Hamid Sharif Nia[1], Erika Sivarajan Froelicher[2,3], Lida Hosseini[4] and Mansoureh Ashghali Farahani[4]**

[1] School of Nursing and Midwifery, Mazandaran University of Medical Sciences, Sari, Iran, [2] Department of Physiological Nursing, School of Nursing, University of California, San Francisco, San Francisco, CA, United States, [3] Department of Epidemiology and Biostatistics, School of Medicine, University of California, San Francisco, San Francisco, CA, United States, [4] School of Nursing and Midwifery, Iran University of Medical Sciences, Tehran, Iran

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Sonja Heintz,
University of Plymouth,
United Kingdom
Irene Checa,
University of Valencia, Spain

***Correspondence:**
Lida Hosseini
l.hosseini69@gmail.com
Mansoureh Ashghali Farahani
m_negar110@yahoo.com

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

Received: 20 December 2021
Accepted: 21 March 2022
Published: 01 June 2022

Citation:
Sharif Nia H, Froelicher ES, Hosseini L and Ashghali Farahani M (2022) Evaluation of Psychometric Properties of Hardiness Scales: A Systematic Review. Front. Psychol. 13:840187. doi: 10.3389/fpsyg.2022.840187

Background: Hardiness is one of the personality traits that can help individuals in stressful situations. Since human beings are constantly under stressful situations and the stresses inflicted on people in each situation are different, various scales have been developed for assessing this feature among different people in different situations. Hence, it becomes necessary for researchers and health workers to assess this concept with valid and reliable scales. This systematic review aims to rigorously assess the methodological quality and psychometric properties of hardiness scales.

Method: In the first step, the databases including Scopus, PubMed, Web of science, and Persian databases were searched using suitable keywords without limitation time. We select eligible suitable studies after screening titles and abstracts. The quality of studies was evaluated using the Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) checklist and the Terwee quality criteria.

Result: Of the 747 articles identified, 33 articles were entered in this study. Based on the COSMIN checklist, the most reported properties were as following structural validity (84%), hypothesis testing (56%), content validity (42%), and internal consistency (39%). Furthermore, 12 studies reported cross-cultural validity, three studies criterion validity, and one study reported measurement error.

Conclusion: The "family caregivers' hardiness scale," "Japanese Athletic Hardiness Scale," "Occupational Hardiness Questionnaire," and "Children's Hardiness Scale" are the best tools for assessing hardiness in family caregivers, athletes, employees, and children respectively. In addition, the "Dispositional Resilience Scale" (DRS-15) and The Personal Views Survey (PVS III-R) are the most frequently used scales with suitable features for measuring hardiness in the general population.

Keywords: hardiness, hardy personality, systematic review, psychometric testing, validation studies, validity, reliability

INTRODUCTION

Human beings are constantly growing and moving from one stage to the other. This personal development process is an unpredictable and demanding process during each of the development stages during stressful circumstances (Maddi, 2004; Sharif Nia et al., 2021). These stress conditions can have a negative effect on performance, motivation, and health if they are not handled well (Bonanno, 2004). It should be noted that in addition to the natural and continuous stresses during the growth process, the current circumstances create conditions that add additional stresses by rapid changes in all spheres of life (Efimova et al., 2019). Many people cannot control these stressful situations. This in turn can threaten the individual's physical, mental, and social aspects of their health (Bigalke, 2015).

Hardiness is one personality trait that can help individuals in stressful situations. The concept of hardiness was first proposed by Kobasa in 1979 based on the existence theory, which is conceptualized as one of the main personality structures for understanding motivation, excitement, and behavior (Kobasa, 1979). This concept finds meaning in the face of stressful situations are considered as a buffered and intervening variable that moderates the relationship between stressful situations and the physical and psychological effects (Abdollahi et al., 2018). Hardiness is a combination of attitudes and beliefs that motivate an individual to do hard and strategic work in the face of stressful and difficult situations (Maddi, 2007). Kobasa defined hardiness as a multidimensional personality trait consisting of three components or the 3C's: commitment, control, and challenge (Kobasa, 1979). Commitment was defined as a tendency to engage in life's activities and to have a genuine interest and curiosity about the world around us (activities, things, and others) and it includes a feeling of personal competence and feeling of community and/or corporation, control was defined as believing and acting as if one can influence the events of one's life, and this belief in influence occurs as part of one's efforts. This feature allows the person to perceive the predictable consequences of their activities in stressful events and manage them favorably (Luceño-Moreno et al., 2020). Finally, the tendency to challenge was defined as the belief that change, rather than stability, as a natural way of life creates opportunities for personal growth rather than a threat to one's security (Kobasa, 1979).

It should be noted that in 2005, Maddi proposed another dimension called connection as the fourth dimension or the 4th C of hardiness (Maddi and Khoshaba, 2005). According to him, individuals gain part of their power and ability to face stressful situations because of communication with other members of society. Therefore, communication is one of the factors that play an important role in creating and maintaining hardiness (Maddi and Khoshaba, 2005). In 2017 Mund proposed culture as the fifth dimension or the 5th C influencing hardiness. In other words, she proposed that hardiness should not be interpreted as a simple approach regardless of culture (Mund, 2017).

Hardiness is a trait that is related to the person and his environment. Because the prevailing social and cultural conditions affect a person's perception and experience of hardship and threat. In addition, his/her understanding of protective factors and how to use them, and through this, the hardiness dimensions and meanings can be formed (Chan, 2000; Benishek et al., 2005; Green et al., 2020). Therefore, by examining this concept in different groups of people with different stressful situations, various definitions, and components of it have been proposed according to the target community and the context and situation of stressful situations (Hosseini et al., 2021). For example, occupational hardiness means endurance and ability in difficult situations and in fact refers to a person's performance based on cognitive assessments (Moreno-Jiménez et al., 2014). Wagnild and Young also conducted studies on the concept of hardiness in older women and concluded that the meaning of this concept in this group of people includes: equanimity, self-efficacy, perseverance, meaningfulness, and existential aloneness (Wagnild and Young, 1988). Likewise, because hardiness can be taught to people, in order to improve this feature and the ability of people to deal with stressful situations and reduce the effects of stress. Different scales have been developed for different groups such as college students, children, nursing students, and managers (Bartone, 1991; Benishek et al., 2005; Moreno-Jiménez et al., 2014). It should be noted that knowing the degree of the hardiness of individuals or evaluating the effectiveness of interventions requires an accurate and valid scale with desirable psychometric properties (Hosseini et al., 2021). Importantly, these scales consist of different dimensions, and some scales do not cover all the dimensions of hardiness. Hence, this systematic review aims to evaluate the psychometric properties of these scales and make recommendations about their use.

METHODS

Study Design

This is a systematic review to evaluate the psychometric properties of the hardiness scales that were conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Moher et al., 2009).

Eligibility Criteria

Eligibility criteria of this study included English and Persian articles describing the psychometric properties of scales/the process of validation/cross-cultural evaluation of the concept of hardiness. Excluded were articles with irrelevant topics, review/systematic review articles, structural equation model or model testing articles, and articles in languages other than Persian and English.

Information Sources

Five electronic databases such as Scopus, PubMed, Science Direct, ProQuest, and Web of Science were searched for English articles. Two Persian databases including Persian SID[1] and MAGIRAN[2] were also searched for Persian articles. Finally, Google Scholar as a search engine and ProQuest database were searched to identify relevant theses. It is noteworthy that

[1]https://www.sid.ir/

[2]http://www.magiran.com/

the reference lists of all identified articles were also searched manually. The search took place from the years 1979–2022.

Search Strategy Electronic

The search strategy was based on the principle that considering a wide range of search terms leads to the best results of related studies. Therefore, in this study, the search strategy was designed taking into account the main concept, which is hardiness, and the type of study, which includes development or psychometric studies and using considering "abstract, title and keywords." These keywords were used: hardiness, hardy personality, personality hardiness validity, validation, reliability, development, and psychometric. The Persian meanings of these keywords were used for searching in Persian databases. It is noteworthy that each database was searched with proper syntaxes (see **Table 1**).

Study Selection

The initial search yielded 747 articles, 77 were from Scopus, 246 were from PubMed, 111 were from Web of Science, 55 were from Science Direct, 169 were from Google Scholar, 47 were from ProQuest, and 42 were from Persian databases. Of the 747 articles initially identified, 33 met all the inclusion criteria. See reasons for exclusion in **Figure 1**.

All of the articles found by searching databases were stored in an EndNote (version X8; Thomson Reuters, New York, NY, United States) file to display duplicate results. Two authors (LH and HN) independently evaluated all articles for inclusion and exclusion. Any discrepancy between the authors was resolved through joint discussions with the third author. See the selection process schematically in **Figure 1**.

Data Extraction

The data were extracted by two researchers (LH and HN) where one was an expert in statistics extracted data and another was an expert in the concept of the study. A data extraction sheet included: first author name, publication year, country, name of scale, target population, face validity, content validity, construct validity (sample size, factor extraction method, rotation methods, selection of the number of factors, name of factors, and total variance), and reliability [consistency: Cronbach's alpha coefficient, stability: Spearman's correlation coefficient, and Intraclass Correlation Coefficient (ICC)] (see **Supplementary Table 1**).

Risk of Bias

The Consensus-based Standards for the selection of health Measurement Instruments (COSMIN) Risk of Bias checklist was used to assess this feature for each of the 33 studies. This tool includes 3 parts with 10 boxes. The first part addresses content validity and includes boxes 1 and 2. This part assesses the relevance and comprehensibility of all items with the target construct and population. Second Part with boxes 3, 4, and 5 addresses internal structure with structural validity, internal consistency, and cross-cultural validity/measurement invariance. The third part with boxes 6, 7, 8, 9, and 10 address the remaining measurement properties including reliability, measurement error, criterion validity, hypotheses testing for construct validity, and responsiveness. The third part focuses on the quality of the (sub)scale as a whole, rather than on item level (Mokkink et al., 2018).

Quality Assessment and Data Analysis

The full text of the articles was evaluated in terms of methodological quality based on the checklist provided by COSMIN. The COSMIN checklist assesses different psychometric properties including: A = internal consistency, B = reliability, C = measurement error, D = content validity, E = structural validity, F = hypothesis testing, G = cross-cultural validity, H = criterion validity and I = responsiveness. Finally, each article was analyzed using a four-point COSMIN score. Each item was classified into four levels including "excellent" as an appropriate methodology, "good" as an adequate level of quality and insufficient relevant information, "fair" as the questionable methodological process, and "poor" as an incorrect methodological process. A methodological quality score per box is obtained by taking the lowest rating of any item in a box ("worst score counts") (Terwee et al., 2012). Finally, Terwee's study criteria were used to analyze the quality criteria of the measured properties (Terwee et al., 2007). The Inter-reviewer reliability was evaluated according to the Cohen's Kappa value. Any discrepancies were resolved through discussion and consensus.

Data Synthesis

Since a general analysis of psychometric properties is not possible, the characteristics of the available articles were used to determine the validity of the instrument.

RESULTS

Study Characteristics

A total of 747 articles were found; of these 42 articles were from the Persian database and 705 articles were from English language databases. Duplicate articles were excluded and 33 articles were reminded and were evaluated using the COSMIN checklist and Terwee study criteria (see PRISMA flow chart, **Figure 1**).

Studies were published between 1986 and 2021; the majority of them were published between 2016 and 2020 (each year $n = 4$). One study was a doctoral thesis (Ferrara, 2019) and the 32 other articles were original and were published in journals. The majority of them were conducted in the United States ($n = 12$) (Funk and Houston, 1987; Pollock and Duffy, 1990; Bartone, 1991; Benishek, 1996; Velasco-Whetsell and Pollock, 1999; Wang, 1999; Benishek and Lopez, 2001; Benishek et al., 2005; Maddi et al., 2006; Madrigal et al., 2016; Weigold et al., 2016) after that Iran ($n = 4$) (Mohsenabadi and Fathi-Ashtiani, 2021; Soheili et al., 2021a,b; Hosseini et al., 2022), Canada ($n = 2$) (McNeil et al., 1986; Lang et al., 2003), Brazil ($n = 1$) (Solano et al., 2016), China ($n = 1$) (Wong et al., 2014), Netherlands ($n = 2$) (Gebhardt et al., 2001; Dymecka et al., 2020), Greece ($n = 1$) (Kamtsios and Karagiannopoulou, 2013), Croatia ($n = 1$) (Kardum et al., 2012),

TABLE 1 | Keywords used in the search for the different databases.

Databases	Search string
PubMed	(((((((validity[Title/Abstract]) OR (validation[Title/Abstract])) OR (reliability[Title/Abstract])) OR ("Factor analysis"[Title/Abstract])) OR (psychometric[Title/Abstract])) OR (development[Title/Abstract])) AND (((hardiness[Title/Abstract]) OR (hardy personality[Title/Abstract])) OR (personality hardiness[Title/Abstract]))) AND (((((((scale[Title/Abstract]) OR (survey[Title/Abstract])) OR (questionnaire[Title/Abstract])) OR (index[Title/Abstract])) OR (Inventor[Title/Abstract])) OR (Test [Title/Abstract])) OR (Measure[Title/Abstract])) OR (Instrument[Title/Abstract])) 187
Scopus	(TITLE-ABS-KEY ["validity" OR "validation" OR "reliability" "development" OR "psychometric"] AND TITLE-ABS-KEY ["hardiness" OR "hardy personality" OR "personality hardiness"]) 77
Web of science	(validity OR validation OR reliability OR "Factor analysis" OR psychometric OR development) AND ("hardiness" OR "hardy personality" OR "personality hardiness")AND (scale OR survey OR questionnaire OR index OR Inventor OR Test OR Measure OR Instrument) 187

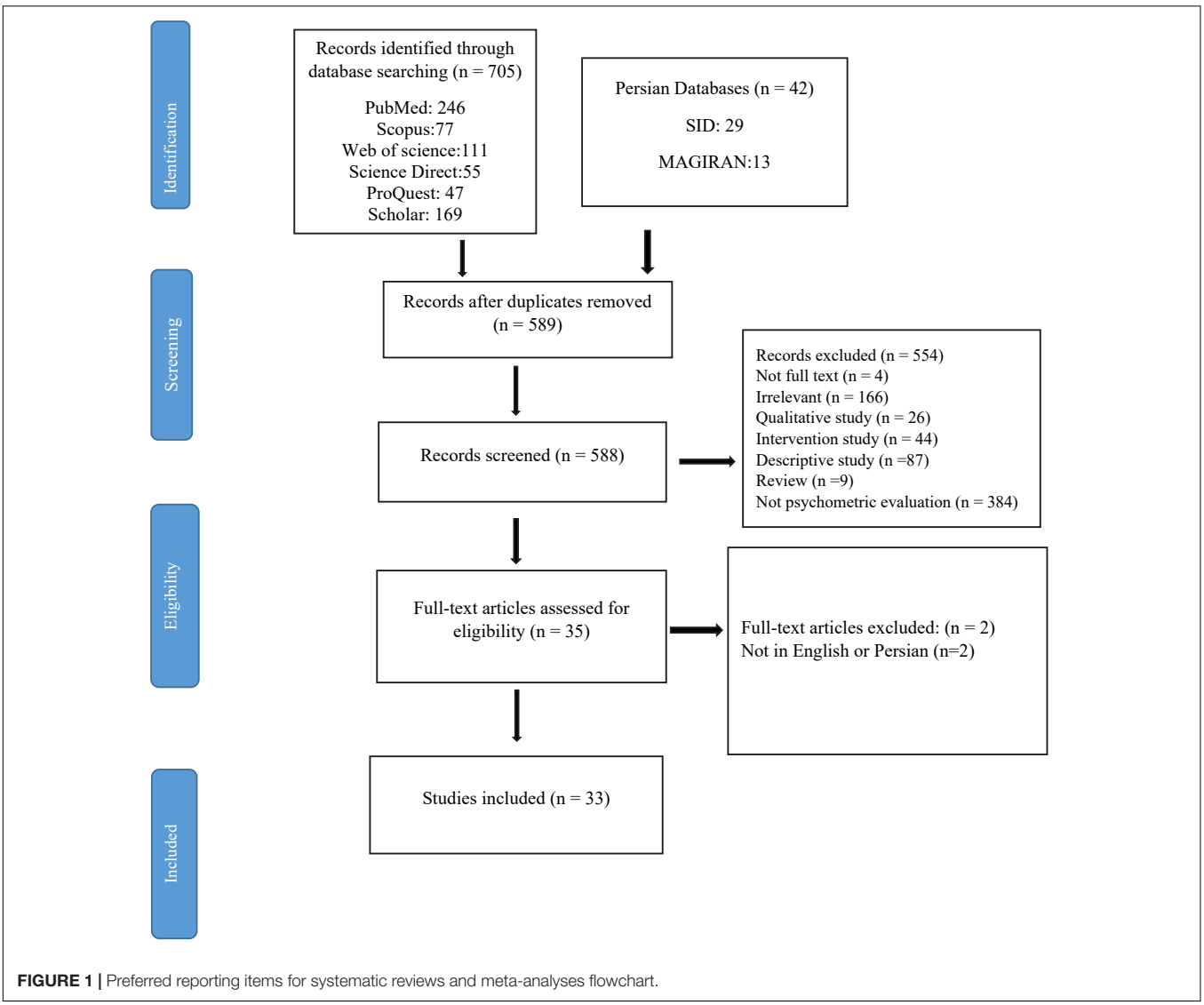

FIGURE 1 | Preferred reporting items for systematic reviews and meta-analyses flowchart.

Italia (n = 1) (Picardi et al., 2012), Spain (n = 2) (Moreno-Jiménez et al., 2014; Luceño-Moreno et al., 2020), Australia (n = 1) (Creed et al., 2013), Sweden (n = 1) (Persson et al., 2016), Taiwan (n = 1) (Cheng et al., 2019), Japan (n = 1) (Yamaguchi et al., 2020), South Korea (n = 1) (Ko et al., 2018), and Norway (n = 1) (Hystad et al., 2010). Only one study was published in the Persian language (Mohsenabadi and Fathi-Ashtiani, 2021). The majority of them were focused on student (n = 10) after that they conducted on general population (n = 6), patients (n = 3), parents (n = 3), military (n = 3), employee (n = 3), health workers (n = 2), athletes (n = 2), and family caregivers (n = 1).

TABLE 2 | The COSMIN risk of bias checklist.

Number	References	BOX 1	BOX 2	BOX 3	BOX 4	BOX 5	BOX 6	BOX 7	BOX 8	BOX 9	BOX10
		PROM development	**Content validity**	**Structural validity**	**Internal consistency**	**Cross-cultural validity \measurement invariance**	**Reliability**	**Measurement error**	**Criterion validity**	**Hypotheses testing for construct validity**	**Responsiveness**
1	McNeil et al., 1986	Very good	Inadequate	Very good	Inadequate	-	Inadequate	Inadequate	-	Very good	
2	Funk and Houston, 1987	Very good	Inadequate	Very good	Inadequate	-	Inadequate	Inadequate	-	Very good	
3	Pollock and Duffy, 1990	Very good	Very good	Very good	Very good	-	Inadequate	Inadequate	-	Adequate	
4	Bartone, 1991	Very good	Inadequate	Doubtful	Adequate	-	Inadequate	Inadequate	-	Inadequate	-
5	Benishek, 1996	Very good	Inadequate	Very good	Doubtful	-	Inadequate	Inadequate	-	Inadequate	-
6	Wang, 1999	Very good	Inadequate	Doubtful	Doubtful	Adequate	Inadequate	Inadequate	-	Inadequate	-
7	Velasco-Whetsell and Pollock, 1999	Very good	Adequate	Inadequate	Inadequate	Adequate	Inadequate	Inadequate	-	Inadequate	-
8	Gebhardt et al., 2001	Very good	Inadequate	Very good	Doubtful	Inadequate	Inadequate	Inadequate	-	Adequate	-
9	Benishek and Lopez, 2001	Very good	Inadequate	Very good	Very good	-	Inadequate	Inadequate	-	Very good	-
10	Lang et al., 2003	Very good	Very good	Very good	Inadequate	Inadequate	Inadequate	Inadequate	-	Very good	-
11	Benishek et al., 2005	Very good	Very good	Very good	Very good	-	Inadequate	Inadequate	-	Very good	-
12	Maddi et al., 2006	Very good	Inadequate	Very good	Doubtful	-	Inadequate	Inadequate	-	Very good	-
13	Hystad et al., 2010	Very good	Inadequate	Very good	Doubtful	-	Inadequate	Inadequate	-	Inadequate	-
14	Kardum et al., 2012	Very good	Inadequate	Very good	Doubtful	-	Inadequate	Inadequate	-	Very good	-
15	Picardi et al., 2012	Very good	Adequate	Doubtful	Doubtful	Adequate	Very good	Inadequate	Doubtful	Inadequate	-
16	Kamtsios and Karagiannopoulou, 2013	Very good	Very good	Very good	Doubtful	-	Very good	Inadequate	-	Inadequate	-
17	Creed et al., 2013	Very good	Adequate	Very good	Inadequate	-	Inadequate	Inadequate	-	Very good	-
18	Moreno-Jiménez et al., 2014	Very good	Very good	Very good	Very good	Inadequate	Inadequate	Inadequate	-	Very good	-
19	Wong et al., 2014	Very good	Very good	Very good	Doubtful	Very good	Inadequate	Inadequate	-	Very good	-

(Continued)

TABLE 2 | (Continued)

Number	References	BOX 1	BOX 2	BOX 3	BOX 4	BOX 5	BOX 6	BOX 7	BOX 8	BOX 9	BOX10
		PROM development	**Content validity**	**Structural validity**	**Internal consistency**	**Cross-cultural validity \measurement invariance**	**Reliability**	**Measurement error**	**Criterion validity**	**Hypotheses testing for construct validity**	**Responsiveness**
20	Persson et al., 2016	Very good	Inadequate	Very good	Doubtful	Adequate	Inadequate	Inadequate	-	Very good	-
21	Weigold et al., 2016	Very good	Inadequate	Very good	Very good	-	Inadequate	Inadequate	-	Very good	-
22	Madrigal et al., 2016	Very good	Inadequate	Very good	Doubtful	-	Inadequate	Inadequate	-	Very good	-
23	Solano et al., 2016	Very good	Very good	Very good	Doubtful	Adequate	Very good	Inadequate	-	Very good	-
24	Ko et al., 2018	Very good	Very good	Very good	Very good	Adequate	Very good	Inadequate	Doubtful	Adequate	-
25	Ferrara, 2019	Very good	Adequate	Doubtful	Doubtful	-	Inadequate	Inadequate	-	Inadequate	-
26	Cheng et al., 2019	Very good	Very good	Very good	Doubtful	-	Inadequate	Inadequate	-	Very good	-
27	Yamaguchi et al., 2020	Very good	Very good	Very good	Very good	-	Inadequate	Inadequate	-	Very good	-
28	Soheili et al., 2021a	Very good	Very good	Very good	Very good	-	Inadequate	Inadequate	-	Very good	-
29	Dymecka et al., 2020	Very good	Inadequate	Very good	Very good	Adequate	Inadequate	Inadequate	Doubtful	Very good	-
30	Luceño-Moreno et al., 2020	Very good	Inadequate	Very good	Very good	-	Inadequate	Inadequate	-	Adequate	-
31	Mohsenabadi and Fathi-Ashtiani, 2021	Very good	Very good	Very good	Very good	Very good	Inadequate	Inadequate	-	Adequate	-
32	Soheili et al., 2021b	Very good	Very good	Very good	Very good	-	Inadequate	Inadequate	-	Adequate	-
33	Hosseini et al., 2022	Very good	Very good	Very good	Very good	-	Very good	Very good	-	-	Very good

Findings From the Risk of Bias Evaluation

Using the COSMIN Risk of bias checklist, the quality of the research manuscripts included in this review was evaluated. From 33 articles, only 45.4% of the studies (15 articles) scored "very good" on both content validity boxes. Also, only 39.3% of the studies (13 articles) scored "very good" on both internal structure boxes. The third part of the risk of bias assessment includes 4 boxes that only 4 studies reported on 3 of 4 boxes as not very good; just one study got a "very good" score in 2 boxes (Hosseini et al., 2022). Details of the risk of bias have been reported in **Table 2**.

Psychometric Properties

Concerning the study design, 15 studies were conducted to develop a scale and 18 of them assessed the psychometric properties. See details of psychometric characteristics in **Supplementary Table 1**. These scales were different based on item number and dimensions. The minimum item number was 12 (Kardum et al., 2012; Dymecka et al., 2020; Yamaguchi et al., 2020) and the maximum was 45 (Lang et al., 2003). Also, the minimum numbers of dimensions were one in two studies (McNeil et al., 1986; Kardum et al., 2012) and one instrument had 9 dimensions (Kamtsios and Karagiannopoulou, 2013). From these 33 studies, 31 studies tested internal consistency, 16 tested test-retest reliability, two studies tested criterion validity (Ko et al., 2018; Dymecka et al., 2020), and 30 studies tested construct validity. Most of the studies evaluated internal consistency and stability using Cronbach's alpha, but four studies evaluated stability using ICC (Picardi et al., 2012; Kamtsios and Karagiannopoulou, 2013; Solano et al., 2016; Hosseini et al., 2022). The criterion validity was tested in two studies (Ko et al., 2018; Dymecka et al., 2020). The construct validity was tested using principal components factor or principal axis factor analysis in most of the studies (n = 16), exploratory factor analysis (n = 3), and confirmatory factor analysis (CFA) was assessed in 10 studies. Five studies did not evaluate the construct validity. The total variance that is explained with these scales ranges from 32.1% to 69% and 15 studies did not report it.

Quality Assessment

The details of the COSMIN quality assessment of 33 articles are shown in **Tables 3**, **4**. None of these articles had "Excellent" quality in all psychometric properties.

BOX A—Internal Consistency

The interrelatedness among the items of each scale was determined by measuring internal consistency. The main quality criteria to evaluate internal consistency are as follows: (1) adequate sample size (seven per items and > 100), (2) calculating Cronbach's alpha (s) for each dimension separately, and (3) Cronbach's alpha (s) between 0.70 and 0.95 (Terwee et al., 2007). Based on these criteria 13 studies were evaluated as "Excellent," one study was "good" because it did not calculate alpha for each dimension/subscale separately (Bartone, 1991). Three studies did not evaluate internal consistency (Funk and Houston, 1987; Velasco-Whetsell and Pollock, 1999; Creed et al., 2013) and were deemed of "poor" quality. Two studies were evaluated as "poor" because did not meet two of the three criteria (McNeil et al., 1986; Lang et al., 2003). Finally, 14 studies were evaluated as "fair" because their Cronbach's alpha (s) were < 0.70 or > 0.95.

BOX B—Reliability

Reliability was used to show that score did not change by repeating the measurement with three methods: (1) test-retest for overtime, (2) inter-rater for measuring by different persons on the same occasion, and (3) intra-rater for measuring by the same persons (i.e., raters or responders) on different occasions. The main quality criteria to evaluate reliability are ICC or weighted Kappa ≥ 0.70 (Terwee et al., 2007). Five studies were evaluated as "Excellent," (Picardi et al., 2012; Kamtsios and Karagiannopoulou, 2013; Solano et al., 2016; Ko et al., 2018; Hosseini et al., 2022), eight studies were evaluated as "poor" because they did not report ICC or Kappa value; and 20 studies did not evaluated reliability and were deemed of "poor" quality.

BOX C—Measurement Error

The means of measurement error is the systematic and random error of a score that cannot be attributed to true changes in the construct reported by the Standard Error of Measurement (SEM). Just one study reported measurement errors (Hosseini et al., 2022).

BOX D—Content Validity

Content validity is defined as "the content of the scale items reflects the structure we intend to measure." The quality criteria to evaluate the content validity are assessment of the relevancy of all items to the construct, the study population, the measurement purpose, and experts involved in item selection. 15 studies did not report content validity and they were evaluated as "poor." Four studies did not mention who was involved in content validity and they were evaluated as "good" (Velasco-Whetsell and Pollock, 1999; Picardi et al., 2012; Creed et al., 2013; Ferrara, 2019) and 14 of others were evaluated as "Excellent."

BOX E—Structural Validity

Structural validity refers to the degree to which the scores obtained from the scale reflect sufficient dimensions of the construct. Main quality criteria that show this feature are performing factor analysis by FEA or CFA. In this review, five studies did not report factor analysis and were evaluated as "fair" (Bartone, 1991; Velasco-Whetsell and Pollock, 1999; Wang, 1999; Picardi et al., 2012; Ferrara, 2019). Other studies were evaluated as "Excellent."

BOX F—Hypothesis Testing

Based on the COSMIN checklist, the purpose of hypothesis testing is the same as construct validity. The main quality criteria that show this feature are formulating specific hypotheses and at least 75% of the results are in accordance with these hypotheses. Nine studies did not report about construct validity and were scored as "poor" (Bartone, 1991; Benishek, 1996;

TABLE 3 | COSMIN quality assessment.

Number	First author (year)	BOX A Internal consistency	BOX B Reliability	BOX C Measurement error	BOX D Content validity	BOX E Structural validity	BOX F Hypothesis testing	BOX G Cross-cultural validity	BOX H Criterion validity
		1. Adequate sample size (≥ 100) 2. Calculate the internal consistency for each dimension (sub)scale 3. Cronbach's alpha (s) between 0.70 and 0.95	1. Available at least two measurements 2. Adequate sample size (≥ 100) 3. Calculated ICC or weighted Kappa ≥ 0.70	Calculated the Standard Error of Measurement (SEM)	Assessment of the relevancy of all items to 1. The construct 2. The study population 3. The measurement Purpose 4. Experts involved in item selection	1. Perform EFA or CFA	1. Specific hypotheses were formulated 2.75% of the results are in accordance with these hypotheses	1. Describing translation process 2. Translating item forward and backward 3. Independently 4. Adequate sample size 5. Pre-testing the scale 6. Performing CFA	1. Using the gold standard 2. Correlation with gold standard is > 0.70
1	McNeil et al., 1986	1. Yes, 2. No, 3. No	1. Yes, 2. Yes, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
2	Funk and Houston, 1987	1. No, 2. No, 3. No	1. No, 2. No, 3. No	No	1. No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
3	Pollock and Duffy, 1990	1. Yes, 2. Yes, 3. Yes,	1. Yes, 2. Yes, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
4	Bartone, 1991	1. Yes, 2. No, 3. Yes	1. Yes, 2. Yes, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
5	Benishek, 1996	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4 . No	Yes	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
6	Wang, 1999	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	No	1. No, 2. No	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. Yes, 6. No	1. No, 2. No
7	Velasco-Whetsell and Pollock, 1999	1. No, 2. No, 3. No	1. No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. No	No	1. No, 2. No	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. Yes, 6. No	1. No, 2. No
8	Gebhardt et al., 2001	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. No	1. No, 2. Yes, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
9	Benishek and Lopez, 2001	1. Yes, 2. Yes, 3. Yes,	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
10	Lang et al., 2003	1. No, 2. Yes, 3. No	1. Yes, 2. Yes, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. Yes, 2. Yes, 3. Yes, 4. No, 5. No, 6. No	1. No, 2. No
11	Benishek et al., 2005	1. Yes, 2. Yes, 3. Yes	1. Yes, 2. Yes, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
12	Maddi et al., 2006	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
13	Hystad et al., 2010	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
14	Kardum et al., 2012	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No

(Continued)

TABLE 3 | (Continued)

Number	First author (year)	BOX A Internal consistency	BOX B Reliability	BOX C Measurement error	BOX D Content validity	BOX E Structural validity	BOX F Hypothesis testing	BOX G Cross-cultural validity	BOX H Criterion validity
15	Picardi et al., 2012	1. Yes, 2. Yes, 3. No	1. Yes, 3. Yes, 4. Yes	No	Yes, 2. Yes, 3. Yes, 4. No	No	1. No, 2. No	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. Yes, 6. No	1. Yes, 2. No
16	Kamtsios and Karagiannopoulou, 2013	1. Yes, 2. Yes, 3. No	1. Yes, 2. Yes, 3. Yes	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
17	Creed et al., 2013	1. No, 2. No, 3. No	1. No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
18	Moreno-Jiménez et al., 2014	1. Yes, 2. Yes, 3. Yes	1. Yes, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. Yes, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
19	Wong et al., 2014	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. Yes, 6. Yes	1. No, 2. No
20	Persson et al., 2016	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. No, 6. Yes	1. No, 2. No
21	Weigold et al., 2016	1. Yes, 2. Yes, 3. Yes	1. Yes, 2. Yes, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
22	Madrigal et al., 2016	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
23	Solano et al., 2016	1. Yes, 2. Yes, 3. No	1. Yes, 2. Yes, 3. Yes	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. Yes, 6. No	1. No, 2. No
24	Ko et al., 2018	1. Yes, 2. Yes, 3. Yes	1. Yes, 2. Yes, 3. Yes	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. No	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. No, 6. Yes	1. Yes, 2. No
25	Ferrara, 2019	1. Yes, 2. Yes, 3. No	1. No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. No	No	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
26	Cheng et al., 2019	1. Yes, 2. Yes, 3. No	No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
27	Yamaguchi et al., 2020	1. Yes, 2. Yes, 3. Yes	No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
28	Soheili et al., 2021a	1. Yes, 2. Yes, 3. Yes	No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. Yes	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
29	Dymecka et al., 2020	1. Yes, 2. Yes, 3. Yes	No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. Yes, 2. Yes	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. No, 6. Yes	1. Yes, 2. No
30	Luceño-Moreno et al., 2020	1. Yes, 2. Yes, 3. Yes	No, 2. No, 3. No	No	No, 2. No, 3. No, 4. No	Yes	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. Yes	1. No, 2. No
31	Mohsenabadi and Fathi-Ashtiani, 2021	1. Yes, 2. Yes, 3. Yes	1. Yes, 2. Yes, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. No, 2. No	1. Yes, 2. Yes, 3. Yes, 4. Yes, 5. Yes, 6. Yes	1. No, 2. No
32	Soheili et al., 2021b	1. Yes, 2. Yes, 3. Yes	1. No, 2. No, 3. No	No	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. Yes, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No
33	Hosseini et al., 2022	1. Yes, 2. Yes, 3. Yes	1. Yes, 2. Yes, 3. Yes	Yes	Yes, 2. Yes, 3. Yes, 4. Yes	Yes	1. No, 2. No	1. No, 2. No, 3. No, 4. No, 5. No, 6. No	1. No, 2. No

TABLE 4 | COSMIN quality assessment.

Number	First author (year)	COSMIN boxes							
		BOX A Internal consistency	**BOX B Reliability**	**BOX C Measurement error**	**BOX D Content validity**	**BOX E Structural validity**	**BOX F Hypothesis testing**	**BOX G Cross-cultural validity**	**BOX H Criterion validity**
1	McNeil et al., 1986	Poor	Poor	Poor	Poor	Excellent	Excellent	-	-
2	Funk and Houston, 1987	Poor	Poor	Poor	Poor	Excellent	Excellent	-	-
3	Pollock and Duffy, 1990	Excellent	Poor	Poor	Excellent	Excellent	Good	-	-
4	Bartone, 1991	Good	Poor	Poor	Poor	Fair	Poor	-	-
5	Benishek, 1996	Fair	Poor	Poor	Poor	Excellent	Poor	-	-
6	Wang, 1999	Fair	Poor	Poor	Poor	Fair	Poor	Good	-
7	Velasco-Whetsell and Pollock, 1999	Poor	Poor	Poor	Good	Fair	Poor	Good	-
8	Gebhardt et al., 2001	Fair	Poor	Poor	Poor	Excellent	Good	Poor	-
9	Benishek and Lopez, 2001	Excellent	Poor	Poor	Poor	Excellent	Excellent	-	-
10	Lang et al., 2003	Poor	Poor	Poor	Excellent	Excellent	Excellent	Poor	-
11	Benishek et al., 2005	Excellent	Poor	Poor	Excellent	Excellent	Excellent	-	-
12	Maddi et al., 2006	Fair	Poor	Poor	Poor	Excellent	Excellent	-	-
13	Hystad et al., 2010	Fair	Poor	Poor	Poor	Excellent	Poor	-	-
14	Kardum et al., 2012	Fair	Poor	Poor	Poor	Excellent	Excellent	-	
15	Picardi et al., 2012	Fair	Excellent	Poor	Good	Fair	Poor	Good	Fair
16	Kamtsios and Karagiannopoulou, 2013	Fair	Excellent	Poor	Excellent	Excellent	Poor	-	-
17	Creed et al., 2013	Poor	Poor	Poor	Good	Excellent	Excellent	-	-
18	Moreno-Jiménez et al., 2014	Excellent	Poor	Poor	Excellent	Excellent	Excellent	Poor	-
19	Wong et al., 2014	Fair	Poor	Poor	Excellent	Excellent	Excellent	Excellent	-
20	Persson et al., 2016	Fair	Poor	Poor	Poor	Excellent	Excellent	Good	-
21	Weigold et al., 2016	Excellent	Poor	Poor	Poor	Excellent	Excellent	-	-
22	Madrigal et al., 2016	Fair	Poor	Poor	Poor	Excellent	Excellent	-	-
23	Solano et al., 2016	Fair	Excellent	Poor	Excellent	Excellent	Excellent	Good	-
24	Ko et al., 2018	Excellent	Excellent	Poor	Excellent	Excellent	Good	Good	Fair
25	Ferrara, 2019	Fair	Poor	Poor	Good	Fair	Poor	-	-
26	Cheng et al., 2019	Fair	Poor	Poor	Excellent	Excellent	Excellent	-	-
27	Yamaguchi et al., 2020	Excellent	Poor	Poor	Excellent	Excellent	Excellent	-	-
28	Soheili et al., 2021a	Excellent	Poor	Poor	Excellent	Excellent	Excellent	-	-
29	Dymecka et al., 2020	Excellent	Poor	Poor	Poor	Excellent	Excellent	Good	Fair
30	Luceño-Moreno et al., 2020	Excellent	Poor	Poor	Poor	Excellent	Good	-	
31	Mohsenabadi and Fathi-Ashtiani, 2021	Excellent	Poor	Poor	Excellent	Excellent	Good	Excellent	-
32	Soheili et al., 2021b	Excellent	Poor	Poor	Excellent	Excellent	Good	-	-
33	Hosseini et al., 2022	Excellent	Excellent	Excellent	Excellent	Excellent	Poor	-	-

Velasco-Whetsell and Pollock, 1999; Wang, 1999; Hystad et al., 2010; Picardi et al., 2012; Kamtsios and Karagiannopoulou, 2013; Ferrara, 2019; Hosseini et al., 2022), six studies did not report enough results and were evaluated as "good" (Pollock and Duffy, 1990; Gebhardt et al., 2001; Ko et al., 2018; Luceño-Moreno et al., 2020; Mohsenabadi and Fathi-Ashtiani, 2021; Soheili et al.,

2021a) and the 18 remaining studies reported construct validity with complete details and were scored as "excellent."

BOX G—Cross-Cultural Validity

According to the COSMIN checklist, cross-cultural research refers to the ability to translate items to reflect the original version of the scale items. The main criteria for assessing these features are as follows: (1) describing the translation process, (2) translating items forward and backward, (3) independently, (4) adequate sample size, (5) pre-testing the scale, and (6) performing Confirmatory Factor Analysis (CFA). Three studies had mentioned that they translated the scale but they did not report the details and were considered "poor" (Gebhardt et al., 2001; Lang et al., 2003; Moreno-Jiménez et al., 2014). Seven studies were evaluated as "good" (Velasco-Whetsell and Pollock, 1999; Wang, 1999; Picardi et al., 2012; Persson et al., 2016; Solano et al., 2016; Ko et al., 2018; Dymecka et al., 2020) because they did not perform CFA or pre-testing. Two studies reported cross-cultural processes with details and they were evaluated as "excellent" (Wong et al., 2014; Mohsenabadi and Fathi-Ashtiani, 2021).

BOX H—Criterion Validity

Criterion validity indicates the degree to which the scores of the scale are an adequate reflection of a "gold standard". The main quality criteria are using the gold standard (having convincing arguments) and the current scale correlates > 0.70 with this gold standard. Three studies had reported the criterion validity as follows: (1) Angelo Picardi et al. performed criterion validity by assessing the correlation between the 15-item Dispositional Resilience Scale (DRS-15) and Psychological Well-Being Scale (as gold standard) (Picardi et al., 2012). (2) Kim et al. reported the criterion validity by assessing correlation among DRS-15, the Korean version of the Center for Epidemiological Studies-Depression Scale (KCES-D), and the Korean Resilience Questionnaire (KRQ-53) (Ko et al., 2018). (3) Dymecka et al. also reported the criterion validity by assessing correlation among health-related hardiness scale (HRHS), Sense of coherence, Self-efficacy, Acceptance of illness, and Psychological resilience (Dymecka et al., 2020). Since the scales that they had chosen were not the gold standard and the correlation between scales was not > 0.70, these studies were evaluated as "fair." It is noteworthy that the responsiveness categories were not analyzed, because there were no results related to that.

DISCUSSION

This study has evaluated the psychometric properties of 33 scales about hardiness using the COSMIN checklist. The salient findings from this study include that no studies have an "Excellent" score for all of the quality criteria of psychometric properties. Therefore, there is no robust and valid single scale for measuring the hardiness concept yet.

This systemic review evaluated all the studies related to psychometric properties about hardiness conducted in different fields, different target populations, different publication times, and countries. Since present life is associated with multiple fast-paced changes and stressful circumstances, individuals in every stage of life, field, and situations need to be able to develop hardiness to face life's difficulties. The results show that the development of scales for hardiness was conducted for any age group from children to older adults. Also, different situations were considered such as students (Benishek and Lopez, 2001; Benishek et al., 2005; Creed et al., 2013; Kamtsios and Karagiannopoulou, 2013; Cheng et al., 2019; Ferrara, 2019; Soheili et al., 2021a), athletes (Yamaguchi et al., 2020), patients (Pollock and Duffy, 1990), general population (McNeil et al., 1986; Funk and Houston, 1987; Bartone, 1991; Maddi et al., 2006; Hystad et al., 2010), parents (Lang et al., 2003; Soheili et al., 2021b), employees (Moreno-Jiménez et al., 2014), and family caregivers (Hosseini et al., 2022). Therefore, some studies were specific for a group of people with a specific situation and some of them were general. As results show, seven scales were developed for students; it may be because students are likely to experience stress and struggle and have had less opportunity to develop hardiness (Cheng et al., 2019). It should be noted that the *Dispositional Resilience Scale (DRS-15)* and *The Personal Views Survey (PVS), PVS II, PVS III, and PVS III-R* are the most frequently used scales and they were translated and assessed in several languages (Hystad et al., 2010; Wong et al., 2014; Madrigal et al., 2016; Solano et al., 2016; Ko et al., 2018; Mohsenabadi and Fathi-Ashtiani, 2021). The newest scale was the "family caregivers' hardiness scale" for family caregivers of patients with Alzheimer's disease (Hosseini et al., 2022).

The dimensions of all scales could be categorized into three themes as designated by Kobasa such as commitment, control, and challenge. Dimension of commitment refers to the tendency toward involvement in the situation as opposed to isolation and explains variances that ranged from 8.92 (Madrigal et al., 2016) to 38.91% (Kamtsios and Karagiannopoulou, 2013) in these studies. The Control dimension refers to belief in the effectiveness of effort on results even in stressful situations. This dimension explains the largest proportion of total explained variance of hardiness in some studies (Pollock and Duffy, 1990; Solano et al., 2016; Yamaguchi et al., 2020). The final dimension is the challenge that refers to perceiving life challenges as a normal part of life and trying to turn them into learning opportunities. This dimension also explains the largest proportion of total explained variance of hardiness in some studies (Hystad et al., 2010; Moreno-Jiménez et al., 2014; Madrigal et al., 2016). The most dimension related to Kamtsios et al. with nine factors of which six factors related to commitment, two factors related to challenging and one factor related to the control dimension (Kamtsios and Karagiannopoulou, 2013).

Since factor extraction uses for raising the explained variance with classifying items into a minimum number of factors, most studies explained total variance $\leq 50\%$; so that the maximum total explained variance is 68.9% related to one study with two factors (Funk and Houston, 1987), and Soheili et al. with 65.75% total variance with three factors (Soheili et al., 2021a). Also, the minimum variance explained according to the study by Pollok (32.1%) reported two factors that measured the effect of hardiness in an individual with an actual health problem (Pollock and Duffy, 1990).

Because the COSMIN checklist is the only standard tool for evaluating the quality of development and psychometric studies. It should be noted that this tool does not report the overall quality scores, because the psychometric properties items are not equal (Terwee et al., 2007). It should be noted that some studies did not report the essential information about psychometric properties clearly and they got a score "poor." Therefore, a low-quality assessment of a scale does not indicate that this scale is inappropriate. In terms of quality, it should also be noted that the quality of more recent articles was better than older publications. This may be due to the development of guidelines by journals for writing and new statistical methods for psychometric evaluation of scales. Another noteworthy point is that most of the studies failed to report face validity, stability, measurement error, and an evaluation of responsiveness, but the newest scale designed in 2022 for family caregivers of patients with Alzheimer's disease has all of these features.

In sum, despite the development of tool guidelines for writing and new statistical methods for psychometric evaluation of scales, each scale has at least one "Poor" psychometric property. Therefore, it is recommended that the COSMIN checklist is used for developing and accessing psychometric properties of scales to provide high-quality scales and future studies should consider features recommended by the COSMIN checklist such as face validity, stability, measurement error, and responsiveness when evaluating the psychometric properties of scales.

Finally based on the results of this systematic review, the highest methodological quality among translation and psychometric studies was the "Korean version of the 15-item Dispositional Resilience Scale" by the Ko et al. study with four boxes of COSMIN checklist scored as "Excellent," two boxes "Good," and one box "Fair" (Ko et al., 2018). Also, the highest methodological quality among development studies was the "family caregivers' hardiness scale" in Hosseini et al. study (Hosseini et al., 2022) with five important boxes of the COSMIN checklist scored as "Excellent," after that the "Occupational Hardiness Questionnaire" in Moreno-Jiménez et al. study (Moreno-Jiménez et al., 2014), "Japanese Athletic Hardiness Scale" in Yamaguchi et al. study (Yamaguchi et al., 2020), and "Children's Hardiness Scale" in Soheili et al. study (Soheili et al., 2021b) with four boxes of COSMIN checklist scored as "Excellent."

Study Limitations

One of the important limitations was lack of access to the full text of the four articles (McCubbin, 1987; Godoy-Izquierdo and Godoy, 2003; Wiedebusch et al., 2007; Grau-Valdes et al., 2020) and lack of assessing two related studies. Because they were in language other than English or Persian (Madrigal et al., 2016; Serrato, 2017).

Study Strength

Hardiness is an important psychological characteristic to deal effectively with stressful situations and reduces the negative physical and psychological effects. Since hardiness can be taught to individuals, knowing which scale has strong validity and reliability characteristics is essential to properly measure this concept. This is the first study that evaluated all scales designed since the introduction of this. Therefore, the findings of this study can help researchers choose the best scale to measure this concept accurately.

Implication

The results of this study can help nurses, researchers, psychologists, health workers, and other decision-makers to identify the best scale concerning quality and psychometric properties.

CONCLUSION

This systematic review provides information about the quality of 33 studies that assessed the psychometric properties of hardiness in various individuals in different stressful situations using the COSMIN checklist. Based on the study results, among developed scales, the "family caregivers' hardiness scale," "Japanese Athletic Hardiness Scale," the "Occupational Hardiness Questionnaire," and "Children's Hardiness Scale" are the best for assessing hardiness in family caregivers, athletes, employees and children. In addition, the Dispositional Resilience Scale (DRS-15) and The Personal Views Survey (PVS III-R) are the most frequently used scales with suitable features for measuring hardiness in the general population.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

LH and HSN designed the study protocol. LH, MA, and HSN searched the data bases and selected the suitable studies. LH and EF wrote the manuscript. All authors approved the final format of manuscript for publication.

ACKNOWLEDGMENTS

We would like to thank all the participants who took part in the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.840187/full#supplementary-material

REFERENCES

Abdollahi, A., Hosseinian, S., Zamanshoar, E., Beh-Pajooh, A., and Carlbring, P. (2018). The moderating effect of hardiness on the relationships between problem-solving skills and perceived stress with suicidal ideation in nursing students. *Studia Psychol.* 60, 30–41. doi: 10.21909/sp.2018.01.750

Bartone, P. T. (1991). Development and validation of a short hardiness measure. *Paper Presented at the Annual Convention of the American Psychological Society*, (Washington, DC: APA).

Benishek, L. A. (1996). Evaluation of the factor structure underlying two measures of hardiness. *Assessment* 3, 423–435. doi: 10.1177/107319119600300408

Benishek, L. A., Feldman, J. M., Shipon, R. W., Mecham, S. D., and Lopez, F. G. (2005). Development and evaluation of the revised academic hardiness scale. *J. Career Assess.* 13, 59–76. doi: 10.1177/1069072704270274

Benishek, L. A., and Lopez, F. G. (2001). Development and initial validation of a measure of academic hardiness. *J. Career Assess.* 9, 333–352. doi: 10.1177/106907270100900402

Bigalke, K. L. (2015). *Coping, Hardiness, and Parental Stress in Parents of Children Diagnosed with Cancer*. Hattiesburg: The University of Southern Mississippi.

Bonanno, G. A. (2004). Have we underestimated the human capacity to thrive after extremely aversive events. *Am. Psychol.* 59, 20–28. doi: 10.1037/0003-066X.59.1.20

Chan, D. W. (2000). Dimensionality of hardiness and its role in the stress-distress relationship among Chinese adolescents in Hong Kong. *J. Youth Adolesc.* 29, 147–161. doi: 10.1023/a:1005100531194

Cheng, Y.-H., Tsai, C.-C., and Liang, J.-C. (2019). Academic hardiness and academic self-efficacy in graduate studies. *High. Educ. Res. Dev.* 38, 907–921. doi: 10.1080/07294360.2019.1612858

Creed, P. A., Conlon, E. G., and Dhaliwal, K. (2013). Revisiting the academic hardiness scale: revision and revalidation. *J. Career Assess.* 21, 537–554. doi: 10.1177/1069072712475285

Dymecka, J., Bidzan-Bluma, I., Bidzan, M., Borucka-Kotwica, A., Atroszko, P., and Bidzan, M. (2020). Validity and reliability of the Polish adaptation of the Health-Related Hardiness Scale – the first confirmatory factor analysis results for a commonly used scale. *Health Psychol. Rep.* 8, 248–262. doi: 10.5114/hpr.2020.95746

Efimova, O. I., Grinenko, A. V., Kalinina, N. V., Miroshkin, D. V., Bazhdanova, Y. V., Oshchepkov, A. A., et al. (2019). Personality hardiness as a factor determining the interaction of a person with the environment (psychological and ecological aspects). *Ekoloji Dergisi* 28, 563–569.

Ferrara, S. (2019). *An Initial Development of a Hardiness Scale for Elementary School Students*. Harrisonburg: James Madison University.

Funk, S. C., and Houston, B. K. (1987). A critical analysis of the Hardiness Scale's validity and utility. *J. Pers. Soc. Psychol.* 53:572. doi: 10.1037/0022-3514.53.3.572

Gebhardt, W., Van der Doef, M., and Paul, L. (2001). The Revised Health Hardiness Inventory (RHHI-24): psychometric properties and relationship with self-reported health and health behavior in two Dutch samples. *Health Educ. Res.* 16, 579–592. doi: 10.1093/her/16.5.579

Godoy-Izquierdo, D., and Godoy, J. (2003). Psychometric properties of the Spanish version of the hardiness scale (personal views survey; PVS). *Psicol. Conductual* 12, 43–78.

Grau-Valdes, Y., Oliva-Hernandez, I., Rojas-Ricardo, L., Grau-Abalo, J. A., and Martinez-Rodriguez, L. (2020). Psychometric properties of the Hardiness Questionnaire (non-work version) in the Cuban population. *Ter. Psicol.* 38, 153–167.

Green, S., Grant, A. M., and Rynsaardt, J. (2020). "Evidence-based life coaching for senior high school students: building hardiness and hope," in *Coaching Researched: A Coaching Psychology Reader*, eds J. Passmore and D. Tee (Hoboken, NJ: John Wiley & Sons), 257–268. doi: 10.1002/9781119656913.ch13

Hosseini, L., Sharif Nia, H., and Farahani, M. A. (2021). Hardiness in family caregivers during caring from persons with Alzheimer's disease: a deductive content analysis Study. *Front. Psychiatry* 12:770717. doi: 10.3389/fpsyt.2021.770717

Hosseini, L., SharifNia, H., and Farahani, M. A. (2022). Development and psychometric evaluation of family caregivers' hardiness scale: a sequential exploratory mixed-method study. *Front. Psychol.* 13:807049. doi: 10.3389/fpsyg.2022.807049

Hystad, S. W., Eid, J., Johnsen, B. H., Laberg, J. C., and Thomas Bartone, P. (2010). Psychometric properties of the revised Norwegian dispositional resilience (hardiness) scale. *Scand. J. Psychol.* 51, 237–245. doi: 10.1111/j.1467-9450.2009.00759.x

Kamtsios, S., and Karagiannopoulou, E. (2013). The development of a questionnaire on academic hardiness for late elementary school children. *Int. J. Educ. Res.* 58, 69–78. doi: 10.5964/ejop.v12i1.997

Kardum, I., Hudek-Knežević, J., and Krapić, N. (2012). The structure of hardiness, its measurement invariance across gender and relationships with personality traits and mental health outcomes. *Psihologijske Teme* 21, 487–507.

Ko, E., Kim, H. Y., Bartone, P. T., and Kang, H. S. (2018). Reliability and validity of the Korean version of the 15-item Dispositional Resilience Scale. *Psychol. Health Med.* 23 (Suppl. 1), 1287–1298. doi: 10.1080/13548506.2017.1417612

Kobasa, S. C. (1979). Stressful life events, personality, and health: an inquiry into hardiness. *J. Pers. Soc. Psychol.* 37:1. doi: 10.1037//0022-3514.37.1.1

Lang, A., Goulet, C., and Amsel, R. (2003). Lang and Goulet hardiness scale: development and testing on bereaved parents following the death of their fetus/infant. *Death Stud.* 27, 851–880. doi: 10.1080/716100345

Luceño-Moreno, L., Talavera-Velasco, B., Jaén-Díaz, M., and Martín-García, J. (2020). Hardy personality assessment: validating the occupational hardiness questionnaire in police officers. *Prof. Psychol. Res. Pract.* 51:297. doi: 10.1037/pro0000285

Maddi, S. R. (2004). Hardiness: an operationalization of existential courage. *J. Hum. Psychol.* 44, 279–298. doi: 10.1177/0022167804266101

Maddi, S. R. (2007). "The story of hardiness: twenty years of theorizing, research, and practice," in *The Praeger Handbook on Stress and Coping*, eds A. Monat, R. S. Lazarus, and G. M. Reevy (Owings Mills, MD: Praeger).

Maddi, S. R., Harvey, R. H., Khoshaba, D. M., Lu, J. L., Persico, M., and Brow, M. (2006). The personality construct of hardiness, III: relationships with repression, innovativeness, authoritarianism, and performance. *J. Pers. Soc. Psychol.* 74, 575–598. doi: 10.1111/j.1467-6494.2006.00385.x

Maddi, S. R., and Khoshaba, D. M. (2005). *Resilience at Work: How to Succeed No Matter What Life Throws at You*. New York, NY: Amacom Books.

Madrigal, L., Gill, D. L., and Eskridge, K. M. (2016). *Examining the Reliability, Validity and Factor Structure of the DRS-15 with College Athletes*. Abingdon: Athletic Performance Research.

McCubbin, M. A. (1987). *Family Hardiness Index*. Madison, WI: University of Wisconsin.

McNeil, K., Kozma, A., Stones, M., and Hannah, E. (1986). Measurement of psychological hardiness in older adults. *Can. J. Aging* 5, 43–48. doi: 10.1017/s0714980800005006

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* 6:e1000097. doi: 10.1371/journal.pmed.1000097

Mohsenabadi, H., and Fathi-Ashtiani, A. (2021). Psychometric properties of the persian version of the dispositional resiliency scale: a brief hardiness measurement scale. *J. Mil. Med.* 23, 338–348.

Mokkink, L. B., De Vet, H. C., Prinsen, C. A., Patrick, D. L., Alonso, J., Bouter, L. M., et al. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual. Life Res.* 27, 1171–1179. doi: 10.1007/s11136-017-1765-4

Moreno-Jiménez, B., Rodríguez-Muñoz, A., Hernández, E. G., and Blanco, L. M. (2014). Desarrollo y validación del Cuestionario Ocupacional de Resistencia. *Psicothema* 26, 207–214. doi: 10.7334/psicothema2013.49

Mund, P. (2017). Hardiness and culture: a study with reference to 3 Cs of Kobasa. *Int. Res. J. Manag. IT Soc. Sci.* 4, 152–159.

Persson, C., Benzein, E., and Årestedt, K. (2016). Assessing family resources: validation of the S wedish version of the F amily H ardiness I ndex. *Scand. J. Caring Sci.* 30, 845–855. doi: 10.1111/scs.12313

Picardi, A., Bartone, P. T., Querci, R., Bitetti, D., Tarsitani, L., Roselli, V., et al. (2012). Development and validation of the Italian version of the 15-item dispositional resilience scale. *Riv. Psichiatr.* 47, 231–237. doi: 10.1708/1128.12446

Pollock, S. E., and Duffy, M. E. (1990). The health-related hardiness scale: development and psychometric analysis. *Nurs. Res.* 39, 218–222.

Serrato, L. (2017). Propiedades psicométricas del cuestionario construido para evaluar personalidad resistente en deportistas (PER-D). *Cuadernos Psicol. Deporte* 17, 25–34.

Soheili, F., Hosseinian, S., and Abdollahi, A. (2021a). Development and initial validation of the children's hardiness scale. *Psychol. Rep.* 124, 1932–1949. doi: 10.1177/0033294120945175

Soheili, F., Hosseinian, S., and Abdollahi, A. (2021b). Development and initial validation of the hardiness based parenting behaviors questionnaire (HBPBQ). *Curr. Psychol.* doi: 10.1007/s12144-021-01673-z

Solano, J. P. C., Bracher, E. S. B., Faisal-Cury, A., Ashmawi, H. A., Carmona, M. J. C., Lotufo, F., et al. (2016). Factor structure and psychometric properties of the Dispositional Resilience Scale among Brazilian adult patients. *Arquivos Neuro Psiquiatr.* 74, 1014–1020. doi: 10.1590/0004-282X20160148

Sharif Nia, H. S., She, L., Rasiah, R., Fomani, F. K., Kaveh, O., Sharif, S. P., et al. (2021). Psychometrics of persian version of the ageism survey among an iranian older adult population during COVID-19 pandemic. *Front. Public Health* 9:683291. doi: 10.3389/fpubh.2021.683291

Terwee, C. B., Bot, S. D., de Boer, M. R., van der Windt, D. A., Knol, D. L., Dekker, J., et al. (2007). Quality criteria were proposed for measurement properties of health status questionnaires. *J. Clin. Epidemiol.* 60, 34–42. doi: 10.1016/j.jclinepi.2006.03.012

Terwee, C. B., Mokkink, L. B., Knol, D. L., Ostelo, R. W., Bouter, L. M., and de Vet, H. C. (2012). Rating the methodological quality in systematic reviews of studies on measurement properties: a scoring system for the COSMIN checklist. *Qual. Life Res.* 21, 651–657. doi: 10.1007/s11136-011-9960-1

Velasco-Whetsell, M., and Pollock, S. E. (1999). The health-related hardiness scale: spanish language equivalence and translation. *Holistic Nurs. Pract.* 13, 35–43. doi: 10.1097/00004650-199904000-00007

Wagnild, G., and Young, H. (1988). *Hardiness among Elderly Women*. San Jose, CA: ERIC.

Wang, J. F. (1999). Verification of the health-related hardiness scale: cross-cultural analysis. *Holistic Nurs. Pract.* 13, 44–52. doi: 10.1097/00004650-199904000-00008

Weigold, I. K., Weigold, A., Kim, S., Drakeford, N. M., and Dykema, S. A. (2016). Assessment of the psychometric properties of the revised academic hardiness scale in college student samples. *Psychol. Assess.* 28:1207. doi: 10.1037/pas0000255

Wiedebusch, S., McCubbin, M., and Muthny, F. (2007). The Family Hardiness Index in German adaptation (FHI-D)-a questionnaire for the assessment of family resiliency. *Pravent. Rehabil.* 19:74. doi: 10.5414/prp19074

Wong, J. Y.-H., Fong, D. Y.-T., Choi, A. W.-M., Chan, C. K.-Y., Tiwari, A., Chan, K. L., et al. (2014). Transcultural and psychometric validation of the Dispositional Resilience Scale (DRS-15) in Chinese adult women. *Qual. Life Res.* 23, 2489–2494. doi: 10.1007/s11136-014-0713-9

Yamaguchi, S., Kawata, Y., Nakamura, M., Murofushi, Y., Hirosawa, M., and Shibata, N. (2020). Development of the revised japanese athletic hardiness scale for University Athletes. *Jpn. J. Appl. Psychol.* 46, 158–166.

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 | Frontiers in Psychology

ORIGINAL RESEARCH
published: 18 July 2022
doi: 10.3389/fpsyg.2022.891959

Typology of Deflation-Corrected Estimators of Reliability

Jari Metsämuuronen [1,2*]

[1] Finnish Education Evaluation Centre, Helsinki, Finland, [2] Centre for Learning Analytics, University of Turku, Turku, Finland

The reliability of a test score is discussed from the viewpoint of underestimation of and, specifically, deflation in estimates or reliability. Many widely used estimators are known to underestimate reliability. Empirical cases have shown that estimates by widely used estimators such as alpha, theta, omega, and rho may be deflated by up to 0.60 units of reliability or even more, with certain types of datasets. The reason for this radical deflation lies in the item–score correlation (*Rit*) embedded in the estimators: because the estimates by *Rit* are deflated when the number of categories in scales are far from each other, as is always the case with item and score, the estimates of reliability are deflated as well. A short-cut method to reach estimates closer to the true magnitude, new types of estimators, and deflation-corrected estimators of reliability (DCERs), are studied in the article. The empirical section is a study on the characteristics of combinations of DCERs formed by different bases for estimators (alpha, theta, omega, and rho), different alternative estimators of correlation as the linking factor between item and the score variable, and different conditions. Based on the simulation, an initial typology of the families of DCERs is presented: some estimators are better with binary items and some with polytomous items; some are better with small sample sizes and some with larger ones.

Keywords: reliability, deflation-corrected reliability, deflation in reliability, coefficient alpha, coefficient theta, coefficient omega, maximal reliability

OPEN ACCESS

Edited by:
Begoña Espejo,
University of Valencia, Spain

Reviewed by:
Rene Gempp,
Diego Portales University, Chile
David Dueber,
University of Kentucky, United States

***Correspondence:**
Jari Metsämuuronen
jari.metsamuuronen@gmail.com

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

Received: *08 March 2022*
Accepted: *30 May 2022*
Published: *18 July 2022*

Citation:
Metsämuuronen J (2022) Typology of Deflation-Corrected Estimators of Reliability. Front. Psychol. 13:891959. doi: 10.3389/fpsyg.2022.891959

INTRODUCTION

From Parallel Test Reliability to Alpha and Maximal Reliability and Beyond From the Perspective of Underestimation in Estimates

> *Reliability has often been underestimated by the conventional formula [. . .]. Many tests are more reliable than they have been considered to be* (Guttman, 1945, p. 260.).

The reliability of a test score generated by a compilation of multiple test items has interested scholars for more than 100 years. In the early phase of the history of measurement modeling, the interest shifted from measurement error to reliability, although measurement error may be a more profound concept than reliability (Gulliksen, 1950). Ever since reliability has become a central measure used to quantify the amount of a random measurement error that exists in a test score. These two concepts are closely linked though because the standard error of the measurement $S.E.m = \sigma_E = \sigma_X\sqrt{1 - REL}$ is defined by reliability $REL = \sigma_T^2/\sigma_X^2 = 1 - \sigma_E^2/\sigma_X^2$ (e.g., Gulliksen, 1950), where σ_T^2, σ_X^2, and refers to the variances of the observed score variable (X), unobserved true score (T), and error element (E) familiar from their profound relation in testing theory,

$X = T + E$. Because the true score T is unobservable, the error element E is also unobservable; therefore, several measurement models based on parallel, tau-equivalent, and congeneric partitions of the test or test items (referring to, e.g., Lord et al., 1968) with different assumptions and multiple estimators of reliability have been developed over the years.

It is well-known that many estimators of reliability underestimate population reliability because of the *attenuation* caused by errors in measurement modeling and random errors in the measurement. However, a less-discussed issue regarding estimates by traditional estimators of reliability is that the estimates may also be radically *deflated* because of artificial systemic errors during the estimation. These concepts are discussed, for instance, by Chan (2008), Lavrakas (2008), Gadermann et al. (2012), Revelle and Condon (2018), and Metsämuuronen (2022a,c,f). Deflation and its correction are the main foci in this article. Some historical turning points and traditional estimators of reliability are discussed from the viewpoint of underestimation in reliability to lead the focus from traditional estimators to the deflation-corrected estimators of reliability discussed in the latter part of the article.

From Brown and Spearman to the Greatest Lower Bound of Reliability

First traces of reliability lead us to Brown (1910) and Spearman (1910), who suggested a way to correct attenuation in the product-moment correlation coefficient (PMC; Bravais, 1844; Pearson, 1896 onward). Pearson (1903) had already noticed that when only a portion of the range of a variable's values is actualized in the sample, this leads to inaccuracy in the estimates of correlation; the estimates are attenuated. This phenomenon is often discussed as range restriction or restriction of range (refer to the literature, e.g., Sackett and Yang, 2000; Sackett et al., 2007; Schmidt et al., 2008; Schmidt and Hunter, 2015). Pearson (1903) and Spearman (1904) were the first to offer solutions to the problem. Later, a coefficient of reliability, the Brown–Spearman prediction formula of reliability based on strictly parallel tests [ρ_{BS}; refer to Cho and Chun (2018) for the history and rationale of the rectified order of innovators], was famously developed to correct the inaccuracy in correlation first by Brown in his unpublished doctoral thesis [before 1910 although referred to in Brown (1910) and later in Spearman (1910). ρ_{BS} is based on a correlation between the strictly parallel partitions g and h of a test. Parallelism implies that the true scores (taus) and variances of a test-taker are assumed to be equal in the sub-tests [$T_g = T_h$, $\sigma_g^2 = \sigma_h^2$; refer to Gulliksen (1950)].

A more useful early innovation based on two partitions, g and h, was offered by Rulon (1939) after being consulted by Flanagan (see the history in Cho and Chun, 2018) based on tau-equivalent partitions: although the lengths of partitions g and h should be equal, they need not be strictly parallel; that is, although the true values of a test-taker are assumed to be (essentially) equal, the variances in the partitions need not be equal ($T_g = T_h$, $\sigma_g^2 \neq \sigma_h^2$). The form of the Flanagan–Rulon prediction formula (ρ_{FR}) appears to be the same as ρ_{BS}, or the form of ρ_{BS} can be expressed in the form of ρ_{FR}(refer to Lord et al., 1968), but the less strict assumptions led to a useful application in the form of the coefficient alpha that will be discussed later. Later, both ρ_{BS} and ρ_{FR}were shown by Guttman (1945) to underestimate population reliability.

Guttman (1945) was the first to show the technical or mechanical basis for underestimation in reliability. All of his six coefficients of reliability ($\lambda_1 - \lambda_6$) were shown to underestimate the true population reliability. Of these, λ_3 and λ_4 appear to be important from the general viewpoint, with λ_4 being a general case of ρ_{BS} and $\rho_{FR}$and λ_3 being equal to the coefficient alpha that will be discussed later. λ_4 was shown to underestimate reliability "*no matter how the test is split*" (Guttman, 1945, p. 260, emphasis original); hence, both ρ_{BS} and ρ_{FR}underestimate the population reliability. The same also applies to an estimator called the greatest lower bound of reliability (ρ_{GLB}) based on λ_4 suggested already by Guttman (1945) and studied later, among others, by Jackson and Agunwamba (1977), Woodhouse and Jackson (1977), and Ten Berge et al. (from Ten Berge and Zegers, 1978 onward; Revelle, 2015; refer also to e.g., Moltner and Revelle, 2015; Trizano-Hermosilla and Alvarado, 2016). Also, McDonald's hierarchical omega ($\rho_{\omega H}$; McDonald, 1999) and Revelle's β (Revelle, 1979; refer also to Zinbarg et al., 2005; Revelle and Zinbarg, 2009) is based on the idea of the *lowest* lower bound of reliability (ρ_{LLB}) belonging to this family [refer to the comparison of estimators based on different types of partition in Revelle (2021) and simulation in Edwards et al. (2021)]. While all the estimators ρ_{BS}, ρ_{FR}, and ρ_{GLB} underestimate the population reliability ($\rho_{population}$), estimators in the framework of ρ_{LLB} give *obvious* underestimations. From the underestimation viewpoint, their relationship is then as follows:

$$\rho_{LLB} < \rho_{FR} \leq \rho_{BS} \leq \rho_{GLB} < \rho_{population}. \quad (1)$$

From Prediction Formulae to Coefficient Alpha

Even before the Flanagan–Rulon formula, Kuder and Richardson (1937) had generalized the idea initiated by Brown and Spearman to a form where each test item in a compilation was taken either as a parallel partition (leading to the coefficient known as KR21, ρ_{KR21}) or a non-parallel although tau-equivalent (or "essentially" tau-equivalent, refer Novick and Lewis, 1967) partition of the test (KR20, ρ_{KR20}). The latter appeared to be more useful in practical testing settings, and it is still in wide use with binary items as one of the lower bounds of reliability.

While KR20 was derived for binary items, the formula was soon generalized to also allow polytomous items (the first usage seems to be in Jackson and Ferguson, 1941; refer to Cho and Chun, 2018), and it was later named coefficient alpha (ρ_α) by Cronbach (1951). Cronbach showed that the estimate by ρ_α is the mean of all split-half partitions (Cronbach, 1951; refer to other interpretations in Cortina, 1993). Warrens (2015) reminds us, though, that this holds only (a) when the partitions include the same number of items, which implies that (b) there are an even number of items on the test to form split-halves with an equal number of items, and (c) when the Flanagan–Rulon formula instead of the Brown–Spearman formula is used.

Because ρ_{KR20}, ρ_{KR21}, and ρ_α are special cases of Guttman's λ_3, they all underestimate the population reliability. Errors in measurement modeling[1] and attenuation have been approximated to cause an underestimation of the magnitude of around 1–11% (see Raykov, 1997a; Graham, 2006; Green and Yang, 2009a; Trizano-Hermosilla and Alvarado, 2016). However, it is generally accepted that when all items are (essentially) tau-equivalent, the phenomenon is unidimensional, and the item-wise errors do not correlate; these estimators would reflect the true reliability (refer to Novick and Lewis, 1967; refer to discussion in, e.g., Cheng et al., 2012; Raykov and Marcoulides, 2017). Unfortunately, this seems to be true only when it comes to *attenuation* in the estimates; this is not true for *deflation,* because the calculation process itself includes a technical or mechanical error that causes deflation in the estimates. The root cause of deflation in ρ_α is the deflation in item–score correlation (ρ_{iX}, *Rit*) embedded in the estimators of reliability; item–score correlation is shown to be severely deflated in settings related to measurement modeling where the scales of the variables deviate radically from each other [refer to algebraic reasons in Metsämuuronen (2016, 2017) and simulations in Metsämuuronen (2020a,b, 2021a, 2022b)]. This element is visible in the form of ρ_α provided in Lord et al. (1968):

$$\rho_\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i\rho_{iX}\right)^2}\right), \quad (2)$$

where k is the number of items in the compilation and σ_i^2 refers to the variance of item g_i. Because of this ρ_{iX}, the estimates of reliability by coefficient alpha may be deflated to the extent of 0.6 units (refer to examples of this magnitude in, e.g., Zumbo et al., 2007; Gadermann et al., 2012; Metsämuuronen and Ukkola, 2019; Metsämuuronen, 2022a,c). Then, from the underestimation viewpoint, the relationship of these estimators is as follows:

$$\rho_{KR20} \leq \rho_{KR21} = \rho_\alpha << \rho_{population}. \quad (3)$$

Despite the known characteristic to underestimate reliability, ρ_α is the most used estimator of reliability in real-life test settings (refer to literature in, e.g., Hoekstra et al., 2019), most probably because of its computational simplicity and obvious conservative nature (e.g., Metsämuuronen, 2017). Because of its wide popularity, alpha has been said to be the *most often wrongly understood* statistic (refer to discussion in, e.g., Sijtsma, 2009; Cho and Kim, 2015; Hoekstra et al., 2019). Therefore, many scholars are ready to remove ρ_α from use (refer to the discussion in, e.g., Sijtsma, 2009; Yang and Green, 2011; Dunn et al., 2013; Trizano-Hermosilla and Alvarado, 2016; McNeish, 2017). However, the issue is still far from settled. Among others, Bentler (2009), Falk and Savalei (2011), Raykov et al. (2014), Metsämuuronen (2017), Raykov and Marcoulides (2017), seem to share stand that when its assumptions are understood and met, ρ_α may be a useful simple tool for assessing (one of) the lower bound(s) of reliability of the score in real-life testing settings. Maybe what is more problematic in the use of ρ_α is that many scholars who use ρ_α may not be able to name *any other* coefficient of reliability that they can use instead. In an empirical study by Hoekstra et al. (2019), 23% of the researchers who published their results in selected renowned journals fell in this group.

From Alpha to Theta, Omega, and Maximal Reliability

The least restricted family of measurement models is based on congeneric partitions of the test. In these models, the true values of the same test-taker need not be identical in the partitions, which means that the assumption of equally long partitions and the same scale in the test items is not required. Also, the weights of items or partitions need not be equal, which allows for multidimensionality in the phenomenon, or the measurement errors, and they need not be independent of each other, too.

Many coefficients of reliability have been developed for these settings. For two congeneric partitions, as counterparts for ρ_{BS} and ρ_{FR}, we have estimators by Angoff and Feldt (ρ_{AF}; Angoff, 1953; Feldt, 1975), Horst (ρ_H; Horst, 1951), and Raju (ρ_β; Raju, 1977). Because the formulae of ρ_{AF} and ρ_β include the same estimate of population variance as in ρ_α: $\sigma_X^2 = \left(\sum_{i=1}^{k}\sigma_i\rho_{iX}\right)^2$, these estimators also tend to give deflated estimates, because the estimate of the item–score correlation by ρ_{iX} is deflated. Based on Warrens (2016), the proportional tendency of these estimators is as follows: if the partitions are equally long, the magnitude of the estimates gets the relationship

$$\rho_{FR} = \rho_\beta \leq \rho_{SB} = \rho_H \leq \rho_{AF} << \rho_{population}, \quad (4)$$

that is, if the condition optimal for ρ_{AF} is fulfilled, other estimators tend to underestimate reliability, and all estimators may produce deflated estimates where the magnitude of the deflation depends on several characteristics such as the difficulty levels of the items. If the variances of the partitions are equal, then

$$\rho_{FR} = \rho_{SB} = \rho_{AF} \leq \rho_H = \rho_\beta << \rho_{population}, \quad (5)$$

that is, if the condition optimal for ρ_H and ρ_β is fulfilled, other estimators tend to underestimate reliability, and all may be radically deflated.

[1] An anonymous reviewer raised the challenge of simplified dimensionality (as part of the error in measurement modeling) as possibly having a profound effect on the underestimation of reliability; if the multidimensionality in the measurement instrument would be considered, the reliability would be profoundly higher (refer to, e.g., McNeish, 2017). From the deflation viewpoint, however, the effect of dimensionality may be less profound, although more studies would enrich the discussion. Namely, even if the multidimensionality would be considered but the items are of extreme difficulty levels in a dimension (as is usual in the achievement testing), the fact remains that the deflation in factor loadings and item score correlations is way more radical than the advance we get from dimensionality. The deflation in factor loadings and item score correlations is discussed in section "PMC as the root cause for the deflation in reliability.

As counterparts to ρ_α for the case in which the scales in items differ from each other, we have two main estimators. For raw scores, we have the Gilmer–Feldt coefficient (ρ_{GF}; Gilmer and Feldt, 1983), also known as the Feldt–Raju coefficient (e.g., Feldt and Brennan, 1989) or the Feldt–Gilmer coefficient (e.g., Kim and Feldt, 2010). Instead of number items (refer to eq. 2), ρ_{GF} uses the proportional weight of the items as a calibrating factor in estimation. The estimates by ρ_α tend to be mildly lower than those by ρ_{GF}. However, the formula of ρ_{GF} uses the same estimate of population variance $\sigma_X^2 = \left(\sum_{i=1}^{k} \sigma_i \rho_{iX}\right)^2$ as ρ_α leading to deflated estimates.

Another alternative for $\rho_\alpha$is to standardize the items and score by principal component analysis (Guttman, 1941), which leads to coefficient theta [ρ_{TH}; Kaiser and Caffrey (1965), based on Lord, 1958], also known as Armor's theta (Armor, 1973). While ρ_α uses raw scores and observed values in items, ρ_{TH} uses standardized items and scores, which has an advantage over ρ_α: the principal component score is one of the "optimal linear combinations" of the score discussed over the years by, chronologically, e.g., Thompson (1940), Guttman (1941), Stouffer (1950), Lord (1958), and Bentler (1968). Zumbo et al. (2007), Gadermann et al. (2012), and Metsämuuronen (2022a,c) have brought ρ_{TH} into discussions again: Zumbo and colleagues because of a new type of estimator called ordinal theta and Metsämuuronen as one of the bases for deflation-corrected estimators of reliability discussed later.

Coefficient theta can be expressed as:

$$\rho_{TH} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} \lambda_{i\theta}^2}\right), \tag{6}$$

where $\lambda_{i\theta}$ is the principal component loadings of the principal component of a one-latent variable model (or of the first principal component), that is, correlations between items and the score variable. It is known that ρ_{TH} maximizes ρ_α (Greene and Carmines, 1980). This can be partly explained by a more effective formula and partly by a more optimally constructed score variable (raw score vs. principal component score). Empirical findings indicate that ρ_{TH} also tends to be conservative (Metsämuuronen, 2022a,f); that is, it seems to underestimate the population reliability although less than the alpha and omega do; the latter will be discussed later. From the viewpoint of underestimation, the relationship of these estimators is then:

$$\rho_\alpha \leq \rho_{GF} < \rho_{TH} < \rho_{population}. \tag{7}$$

In the recent decades, much effort has been gone to explore different aspects of estimators of reliability within the framework of factor models or, more generally, within the latent variable modeling (of the models, refer to, e.g., McDonald, 1985, 1999; Raykov and Marcoulides, 2010). Two of the most discussed estimators are coefficient omega total (ρ_ω; later, just omega), based on the studies of Heise and Bohrnstedt (1970) and McDonald (1970, 1999), and coefficient rho or maximal reliability (ρ_{MAX}; for instance, Raykov, 1997b, 2004), also known as Raykov's rho (refer to, e.g., Cleff, 2019) and Hancock's H (Hancock and Mueller, 2001), based on the conceptualization of "optimal linear combination" discussed above, and later unified by Li et al. (1996) and Li (1997). The two estimators are based on conventions related to factor analysis and factor loadings ($\lambda_{i\theta}$). An ancestor of this family is ρ_{TH}, which is based on the principal component analysis discussed above.

Coefficient omega can be expressed as follows:

$$\rho_\omega = \frac{\left(\sum_{i=1}^{k} \lambda_{i\,\theta}\right)^2}{\left(\sum_{i=1}^{k} \lambda_{i\,\theta}\right)^2 + \sum_{i=1}^{k}\left(1 - \lambda_{i\,\theta}^2\right)}, \tag{8}$$

and rho as:

$$\rho_{MAX} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k} \left(\lambda_{i\theta}^2 / \left(1-\lambda_{i\theta}^2\right)\right)}}, \tag{9}$$

where $\lambda_{i\theta}$ refers to factor loadings by maximum likelihood estimation of a one-latent variable model, although models with multiple dimensions are also in use. The measurement model related to these estimators will be discussed later.

In the theoretical case where all item weights are equal, ρ_{TH}, ρ_ω, and ρ_{MAX} are equal to ρ_α. From this viewpoint, it may be correct to conclude that ρ_{TH}, ρ_ω, and ρ_{MAX} are general forms of ρ_α (refer to, e.g., Hayes and Coutts, 2020). Otherwise, the magnitude of the estimates by ρ_α is smaller than by ρ_{TH} (Greene and Carmines, 1980), and the magnitude of the estimates by ρ_ω is smaller than by ρ_{MAX} (e.g., Cheng et al., 2012). Hence, it seems that both ρ_α and ρ_ω tend to underestimate reliability. A possible confounding phenomenon is that the estimates of reliability by ρ_{MAX} tend to be overestimated with finite or small sample sizes (refer to Aquirre-Urreta et al., 2019; Metsämuuronen, 2022a,c,f). This is caused by the fact that even if only one item has loading $\lambda_i \approx 1$, the element $\lambda_i^2/\left(1-\lambda_i^2\right)$ in eq. (9) becomes unstable and gives, most probably, a value too high compared to the population. This may happen easily with small sample sizes because they are prone to produce deterministic or near-deterministic patterns of the item–score relationship (see discussion in Metsämuuronen, 2022c,f). From the viewpoint of underestimation, in practical settings excluding the theoretical case of identical factor loadings, the relationship of these estimators is then:

$$\rho_\alpha < \rho_\omega < \rho_{TH} < \rho_{MAX} < \rho_{population}\,(< \rho_{MAX})\,. \tag{10}$$

In real-life settings, the difference between the estimates by ρ_α, ρ_{TH}, ρ_ω, and ρ_{MAX} may be subtle. For example, in a simulation

with 1,440 real-life datasets (Metsämuuronen, 2022f), the average magnitude of the lowest estimates by ρ_α was 0.024 units of reliability (2.9%) lower than the highest estimates by ρ_{MAX}. Similarly, the average estimate byρ_ω was 0.021 units (2.4 %) lower than by ρ_{MAX} and 0.017 units (1.9 %) lower than by ρ_{TH}. Notably, though, the difference between ρ_α and ρ_{MAX} seems to be the wider the smaller the sample size is. In the simulation (Metsämuuronen, 2022f), with a sample size of $n = 25$, the average difference between ρ_α and ρ_{MAX} was 0.056 units of reliability (6.4 %), and with $n = 200$, the difference was just 0.008 units of reliability (0.92 %).

From Alpha, Theta, Omega, and Rho to Deflation-Corrected Reliability

While ρ_α is known to underestimate reliability, it seems that ρ_{TH}, ρ_ω, and ρ_{MAX} also tend to give obvious underestimates with certain kinds of datasets, typically with tests of extreme difficulty levels or with incremental difficulty levels including both very easy and very difficult test items. This is a reasonable conclusion from the known character of PMC embedded in the traditional estimators of reliability in the form of *Rit* and λ_i to underestimate the true correlation when the scales of two variables are far from each other as is typical with an item and the score variable (e.g., Metsämuuronen, 2022a,c,f; refer later to **Figure 1**). Recall the relationship between PMC $= \rho_{gX} = Rit$ and the principal component loading (in ρ_{TH}) and factor loading (in ρ_ω and ρ_{MAX}): the loading λ_i is, essentially, a correlation between an item and a score variable (e.g., Cramer and Howitt, 2004; Yang, 2010).

Knowing that PMC is always deflated in cases where scales in the variables are not equal, as is always the case between an item and the score variable, all the estimators mentioned above are deflated, sometimes radically. Empirical findings show that the estimates by ρ_α, ρ_{TH}, ρ_ω, and ρ_{MAX}may be deflated by 0.4–0.6 units of reliability or 46–71% as discussed above (refer to examples in, e.g., Zumbo et al., 2007; Gadermann et al., 2012; Metsämuuronen and Ukkola, 2019; Metsämuuronen, 2022a,c,f). Metsämuuronen (2022a) notes that deflation of this size is remarkable and needs to be studied because it is no more caused by an error in the measurement modeling such as violations in tau-equivalency, unidimensionality, or uncorrelated errors as is traditionally suggested (refer to above). From this point of view, the deflation of 0.4–0.6 units of reliability must be explained directly by some mechanical reasons, and this raises the issue of underestimation in reliability to a new level.

Metsämuuronen (e.g., 2022a; 2022b; 2022f) has used the concept of "mechanical error in the estimates of correlation" (MEC) to understand deflation. The obvious and grave deflation in traditional estimators of reliability has motivated the development of and studies on new types of estimators of reliability called MEC-corrected estimators of reliability (MCERs; Metsämuuronen, 2022a,f) and attenuation-corrected estimators of reliability (ACERs, Metsämuuronen, 2022c), which are both called deflation-corrected estimators of reliability (DCERs; Metsämuuronen, 2022a,f). In MCERs, the embedded *Rit* and λ_i are replaced by totally *different* estimators of correlation, while in ACERs, *Rit* and λ_i are replaced by *attenuation-corrected* estimators of correlation. The logic for and forms of these estimators are discussed in Metsämuuronen (2022a), and these will be briefly discussed later. Notably, the ordinal alpha and ordinal theta by Zumbo et al. (2007; refer also to Gadermann et al., 2012) may be included as part of the extended family of DCERs, as, instead of changing the item–score correlation itself, the inter-item matrices of PMCs are replaced by matrices of polychoric correlation coefficients.

From the attenuation and deflation viewpoint, in general, the relationship of these estimators is

$$\rho_\alpha < \rho_\omega < \rho_{TH} < \rho_{MAX} << \rho_{DCER} < \rho_{population}. \quad (11)$$

Notably, though, certain DCERs based on rho may be prone to overestimating the population reliability with small sample sizes, because rho itself tends to overestimate reliability with small sample sizes (refer to Aquirre-Urreta et al., 2019), while other DCERs based on alpha, theta, and omega, as being more conservative, may be prone to underestimation (see Metsämuuronen, 2022f). This area is largely unstudied, and the current study intends to shed some light on this issue.

Except for the more established coefficient by Zumbo et al. (2007), studies concerning estimators from the family of DCERs are either at a very initial stage (e.g., Metsämuuronen, 2016, 2018), or they give some examples only of the new possibilities (Metsämuuronen, 2020a,b, 2021a,b, 2022b), or they are based on small example datasets and are fragmentary (refer to Metsämuuronen, 2022a,c,f). The simulations by Metsämuuronen (2022c,f) included a limited comparison of the behavior of some DCERs in comparison with the traditional counterparts using 1,440 estimates based on real-life datasets. This study is intended to give more systematic information on these new estimators by comparing their characteristics under different conditions.

Research Questions

Different families of DCERs can be classified by estimators used as the base (e.g., ρ_α, ρ_{TH},ρ_ω, and ρ_{MAX}, discussed above), by the score variables (e.g., θ_X, θ_{PC}, θ_{FA}, θ_{IRT}, and $\theta_{Non-Linear}$, discussed below), and by the weighting factors between the item and the score variable (e.g., R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, and E_{AC}, discussed below). Combinations are, therefore, many. Systematic studies on the behavior of different combinations would, first, enrich our knowledge of the entire phenomenon and, second, help us to typologize the estimators: which estimators would suit which conditions.

The aim of this study is, first, to compare the characteristics of different DCERs and to form a typology of the estimators: under which conditions which coefficient would be the best option? Second, which combinations of the base and weight factor tend to produce under- or overestimates of reliability in real-life testing settings? In the empirical section, the traditional estimators, alpha, theta, omega, and rho, are used as benchmarks and estimated using their traditional score variables (θ_X, θ_{PC}, and θ_{FA}), while DCERs are restricted to the raw score (θ_X).

Before the empirical section, some elementary conceptual points are discussed briefly to make the notation of DCERs understandable. First, the main reason for deflation in reliability, PMC imbedded in the traditional estimators of reliability, is discussed. Second, the traditional model without the elements related to deflation and a general model including these elements are discussed. Finally, different theoretical bases for DCERs related to the forms of ρ_{α}, ρ_{TH},ρ_{ω}, and ρ_{MAX} are briefly discussed (for more details, refer to, e.g., Metsämuuronen, 2022a,c).

CONCEPTUAL AND OPERATIONAL BASES FOR DCERS

PMC as the Root Cause of Deflation in Reliability

The reason for technical and mechanical deflation in reliability is that traditional estimators of reliability embed PMC in the form of *Rit* and λ_i. PMC is known to be seriously affected by many sources of mechanical error when the scales of two variables are far from each other as is always the case with item and score. In simulations (Metsämuuronen, 2021a, 2022b), seven sources of MEC caused cumulative negative bias in PMC. The sources include extreme item difficulty, a small number of categories in the item, large number of tied cases in the score, and a normally distributed score instead of uniformly distributed. Then, as an example, if the test items are few (leading to a score with a narrow scale with a high number of tied cases), they have an extreme level of difficulty and a binary scale, and the score is normally distributed, we would expect to have radically more deflated item-total correlations leading to radically deflated estimates of reliability, than if the test items are many, they have an average difficulty level, their scale is wide if not continuous, and the score is evenly distributed without tied cases. Notably, this has obvious relevance to the estimates of reliability: If the score does not include tied cases, i.e., because of being continuous or the number of test-takers is small, we expect less deflation in reliability compared with the case that we have a normally distributed or skewed score. However, the effect of skewness in distribution is far less notable than the effect of item difficulty (refer to Metsämuuronen, 2022b, Appendix 1 in **Supplementary Material**; also, refer later to footnote 4). The issue of the effect of the item distribution is further discussed by Olvera Astivia et al. (2020) and the effect of the scale distribution by Foster (2021) and Xiao and Hau (2022).

Several alternatives for *Rit* and λ_i are studied from the viewpoint of technical or mechanical errors in the estimates. To some extent, the MEC-affected behavior is known for such traditional estimators of correlation as polychoric correlation coefficient (RPC; Pearson, 1900, 1913; refer to simulations in Metsämuuronen, 2020a,b, 2021a, 2022b), biserial (R_{BS}) and polyserial correlation (R_{PS}) coefficients (Pearson, 1909; see Metsämuuronen, 2020a), r-bireg and r-polyreg correlation (RREG; Livingston and Dorans, 2004; Moses, 2017; refer to Metsämuuronen, 2022b), item–rest correlation (Rir; Henrysson, 1963; refer to Metsämuuronen, 2018, 2021a), lambda and tau (Goodman and Kruskal, 1954; refer to Metsämuuronen, 2020a), coefficient eta (Pearson, 1903, 1905; refer to Metsämuuronen, 2020a, 2022d), delta (D; Somers, 1962; refer to Metsämuuronen, 2020a,b, 2021a,b, 2022b), gamma (G; Goodman and Kruskal, 1954; refer to Metsämuuronen, 2021a,b, 2022b), and tau-a and tau-b (Kendall, 1938, 1948; refer to Metsämuuronen, 2021b, 2022b). Also, some new estimators are developed and studied from this perspective: generalized discrimination index (GDI, Metsämuuronen, 2020c; also refer to the visualization in Metsämuuronen, 2022e) based on Kelley's discrimination index (Kelley, 1939), dimension-corrected *D* (D2; Metsämuuronen, 2020b, 2021a; refer to simulations in Metsämuuronen, 2021a, 2022b), dimension-corrected *G* (G2; Metsämuuronen, 2021a; refer to simulations in Metsämuuronen, 2021a, 2022b), attenuation-corrected *Rit* (R_{AC}; Metsämuuronen, 2022c,d; refer to simulation in Metsämuuronen, 2022b), and attenuation-corrected eta (E_{AC}; Metsämuuronen, 2022d; refer to a simulation in 2022b).

Of the coefficients of correlation, R_{PC} and R_{REG} reflect a correlation between unobservable *theoretical* constructs, which may be problematic from the testing theory viewpoint (refer to the critique by Chalmers, 2017); we do not have access to these theoretical constructs. From this viewpoint, such estimators of correlation as *G* and *D* reflect an association between two *observed* constructs; in the settings of measurement modeling, and they strictly indicate the proportion of logically ordered test-takers in a test item after they are ordered by the score (refer to Metsämuuronen, 2021b). For example, if *D* is 0.7, 85% of the observations are logically ordered in the ascending order in the item after they are ordered by the score ($p = 0.5 \times 0.70 + 0.5 = 0.85$; refer to Metsämuuronen, 2021b). Because of their conservative nature, with polytomous items having more than three categories, Metsämuuronen (2021a) suggests using *G* and *D* with binary items and with polytomous items having less than four categories. Dimension-corrected *G* and *D* (G_2 and D_2) with semi-trigonometric nature can be used for binary and polytomous items, and in a binary case, $G = G2$ and $D = D_2$. Of the attenuation-corrected estimators of correlation (R_{AC} and E_{AC}), R_{AC} is more conservative than E_{AC}. This follows strictly from the behavior of *Rit* and coefficient *eta*: except for the binary case, where *Rit* and *eta* give identical estimates, the estimates by E_{AC} tend to be higher than those by R_{AC} (refer to Metsämuuronen, 2022d).

The phenomenon of mechanical error in the estimators of correlation is easy to illustrate using two identical (latent) variables with an obvious perfect (latent) correlation ($R = 1$). Let us take the vector of $n = 1{,}000$ normally distributed cases and double it. Of these identical variables with (obvious) perfect correlation, one (item *g* to be) is divided into four categories [0–3; $df(g) = 3$] with difficulty level $p(g) = 0.2$ and the other (score *X*) is divided into 61 categories [0–60; $df(X) = 60$] with an average difficulty level of $p(X) = 0.5$. The difference between the latent correlation and the observed correlation indicates strictly the magnitude of MEC in the estimates (**Figure 1**). Notably, the estimates by such known estimators of the item–score correlation as *tau-b*, *Rir*, *Rit*, *eta*, and Spearman rank-order correlation cannot reach the latent perfect correlation but, instead, include a remarkable magnitude of deflation (>

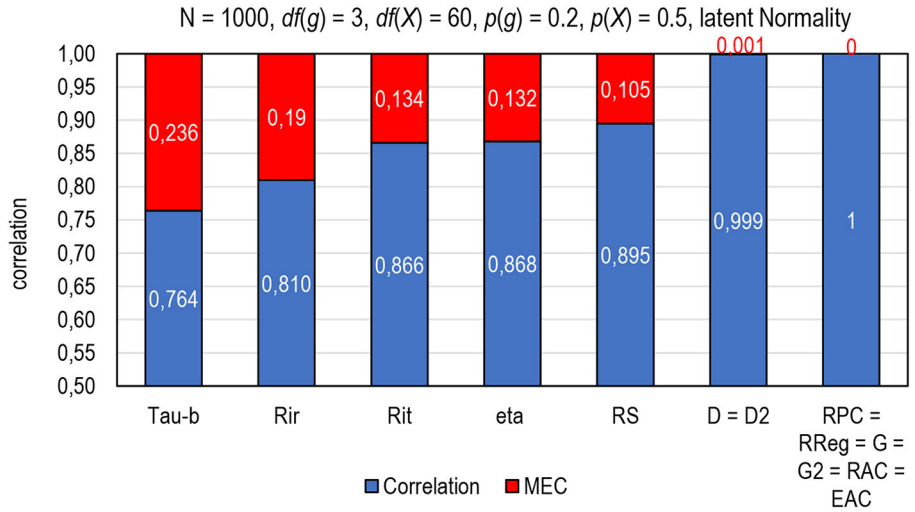

FIGURE 1 | The magnitude of a mechanical error on the estimates of correlation (MEC) by selected estimators of correlation. Tau-b = Kendall tau-b; Rir = Henrysson item–rest correlation (= PMC), Rit = item–total correlation (= PMC); eta = coefficient eta (X dependent), RS = Spearman rank-order correlation (= PMC), D = Somers delta (X-dependent); D2 = dimension-corrected D; RPC = polychoric correlation; R_{REG} = r-polyreg correlation; G = Goodman-Kruskal gamma; G2 = dimension-corrected G, RAC = attenuation-corrected Rit, EAC = attenuation-corrected eta.

0.1 units of correlation) caused by technical and mechanical errors in the estimates. On the contrary, such estimators as R_{PC}, R_{REG}, G, G_2, R_{AC}, and E_{AC} are found MEC-*free* in several conditions (Metsämuuronen, 2022b), and in D and D_2, the magnitude of MEC may be nominal depending on the number of tied pairs in the items and score as well as widths of the scales in the items and score (refer to Metsämuuronen, 2021a).

General Measurement Model Without MEC

Assume a general, simplified, one-latent variable measurement model combining the observed values of an item g_i (x_i), a latent variable (θ), and a weight factor, w_i,that links θ with x_i:

$$x_i = w_i\,\theta + e_i, \tag{12}$$

(e.g., Metsämuuronen, 2022a,c) generalized from the traditional model (e.g., McDonald, 1999; Cheng et al., 2012). In the general model, the theoretical, unobservable θ may be manifested as a varying type of relevantly formed compilation of items including a raw score (θ_X), a principal component score (θ_{PC}), a factor score (θ_{FA}), a theta score formed by the item response theory (IRT) or Rasch modeling (θ_{IRT}), or various non-linear combinations of the items ($\theta_{Non-Linear}$). In the general model, the weight factor w_i is a coefficient of correlation in some form that also includes principal components and factor loadings (λ_i). In all cases, $-1 \leq w_i \leq +1$.

From the coefficient of correlation viewpoint, such estimators as R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, and E_{AC} have been found to be notably better options than PMC (Metsämuuronen, 2022b) as discussed above. In a comparison of eleven sources of MEC, the rough order of the magnitude of MEC ($e_{wi\,\theta\_MEC}$; "MEC" in **Figure 1**) was $e_{PMCi\theta\_MEC} >> e_{Di\theta\_MEC} > e_{D2i\theta\_MEC} >> e_{RREGi\theta\_MEC} > e_{RPCi\theta\_MEC} \approx e_{Gi\theta\_MEC} \approx e_{G2i\theta\_MEC} \approx e_{RACi\theta\_MEC} \approx e_{EACi\theta\_MEC} \approx 0$ (Metsämuuronen, 2022b). That is, of the better behaving estimators above, on the one hand, D is the most conservative option followed by D_2, because both are affected by the number of tied cases in the score variable (refer to Metsämuuronen, 2020b, 2021b). G and D tend to give obvious underestimates with polytomous items with more than 3–4 categories in the scale, so, G_2 and D_2 are suggested to be used with polytomous items instead of G and D (Metsämuuronen, 2021a). On the other hand, using G and D gives quite interesting benchmarking interpretations for the estimates of reliability. Because G and D strictly indicate the proportion of the logically ordered test-takers in a test item after they are ordered by the score ($p = 0.5 \times G + 0.5$ and $p = 0.5 \times D + 0.5$; refer to Metsämuuronen, 2021b), when $D = 0.8$, 90% of the test takers' item responses are in a logical order after the test-takers are ordered by the score. Then, an estimator of reliability using G or D reflects the proportion of logically ordered test-takers in the entire set of test items.

Notably, the estimates by *eta* and *Rit* are identical with binary items; hence, R_{AC} and E_{AC} are identical in binary settings (Metsämuuronen, 2022d). Also, in real-life settings, the sample estimates by R_{AC} and E_{AC} tend to mildly overestimate the populations of R_{AC} and E_{AC} with polytomous items (Metsämuuronen, 2022c,d). This is caused by the fact that a large population rarely includes deterministic patterns between two variables. Hence, the magnitude of the population values of R_{AC} and E_{AC} tend to be somewhat lower than those by sample estimates.

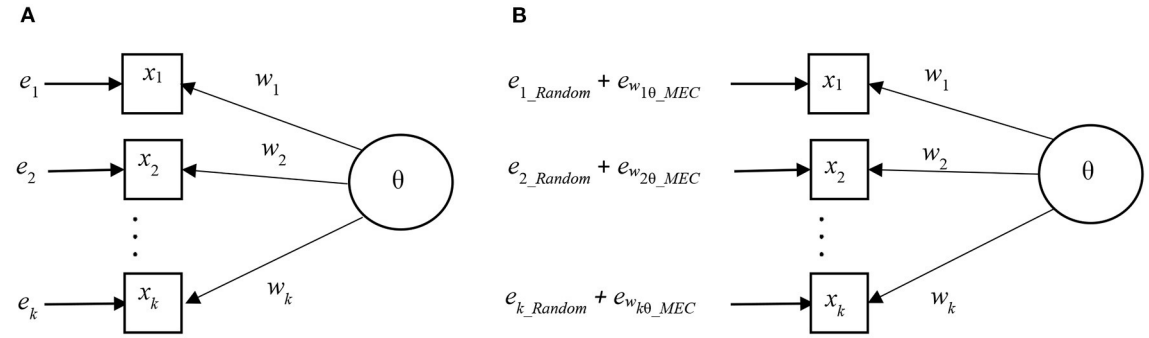

FIGURE 2 | Measurement models without and with elements of MEC. **(A)** Traditional measurement model. **(B)** Measurement model including elements of MEC.

All generally used estimators of correlation give an identical estimate of the correlation for original variables (g_i and θ) and standardized forms of the variables [std(g_i) and std(θ)]. Hence, without loss of generality, to lead to a simple form of the estimators of reliability, let us assume that both item (g_i) and the manifestation of the latent variable (θ) are standardized, that is, $x_i,\ \theta \tilde{}\ N(0,1)$. Then, the item-wise error variance ψ_i^2 is:

$$\psi_i^2 = 1 - w_i^2. \tag{13}$$

From eq. (11), the sum of items is:

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} w_i\,\theta + \sum_{i=1}^{k} e_i, \tag{14}$$

where the error variance related to the compilation of the items is:

$$\sigma_E^2 = \sum_{i=1}^{k} \psi_i^2 = \sum_{i=1}^{k} \left(1 - w_i^2\right), \tag{15}$$

which can be used in estimating the reliability of the score. If θ is manifested as raw score and w_i as *Rit*, eq. (15) could be used in calculating alpha (Eq. 2), although the practicalities lead to the use of different operationalization of the measurement model. If θ is manifested as a principal component score variable and w_i as principal component loadings, the model in eq. (15) leads to theta (eq. 6). If θ is manifested as a factor score variable and w_i as factor loadings, the model in eq. (15) leads to omega and rho (eqs. 8 and 9, respectively).

General Measurement Model Including Elements Related to MEC

The traditional measurement model related to the estimators of reliability assumes that *Rit* and factor/principal component loadings are deflation-free. This is a too optimistic assumption, as illustrated in **Figure 1**. Knowing that a certain part of the measurement error is strictly technical or mechanical but that its magnitude could be reduced, Metsämuuronen (2022a,c) suggested reconceptualizing the classic relationship of $X = T + E$as:

$$X = T + \left(E_{Random} + E_{MEC}\right), \tag{16}$$

where the element E_{MEC} related to deflation is visible. Consequently, we can reconceptualize the measurement model in eq. (12) as:

$$x_i = w_i \times\ \theta\ + \left(e_{i\_Random} + e_{wi\,\theta\,\_MEC}\right), \tag{17}$$

where the element $e_{wi\,\theta\_MEC}$ refers to the fact that the magnitude of the mechanical error depends on the characteristics of the weighting factor w, item i, and score variable θ. In visual forms, the traditional and the MEC-including measurement models are illustrated in **Figures 2A,B** (Metsämuuronen, 2022a). Notably, in **Figure 2**, the magnitude of the error in both models is equal, but in **Figure 2B**, the elements related to MEC are visible.

If we select a weight factor w_i such that the magnitude of the mechanical error is as small as possible, the magnitude of the error component related to deflation may be near zero, that is, $e_{wi\,\theta\,\_MEC} \approx 0$. This would lead to an MEC-corrected (MECC) measurement model where the measurement error would be as near the MEC-free condition as possible, that is:

$$\begin{aligned} x_i &= w_{i\_MECC} \times \theta\ + \left(e_{i\_Random} + e_{wi\theta\,\_MEC}\right) \\ &\approx w_{i\_MECC} \times \theta\ + e_{i\_Random} \end{aligned}. \tag{18}$$

The measurement model with a near-MEC-free weight factor such as R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, and E_{AC}, is illustrated in **Figure 3**.

This conceptualization leads to item-wise MEC-corrected error variance ($\psi_{i\_MECC}^2$):

$$\sigma_{E\_MECC}^2 = \psi_{i\_MECC}^2 = 1 - w_{i\_MECC}^2, \tag{19}$$

where $e_{i\_MECC} \tilde{}\ N\left(0, \psi_{i\_MECC}^2\right)$ and $\psi_{i\_MECC}^2 = 1 - w_{i\_MECC}^2$. Then, after MEC-correction, eq. (15) can be written as:

$$\sum_{i=1}^{k} x_i = \sum_{i=1}^{k} w_{i\_MECC} \times \theta\ + \sum_{i=1}^{k} e_{i\_Random}, \tag{20}$$

and the MEC-corrected error variance of the test score can be written as:

$$\sum_{i=1}^{k} \psi_{i\_MECC}^2 = \sum_{i=1}^{k} \left(1 - w_{i\_MECC}^2\right), \tag{21}$$

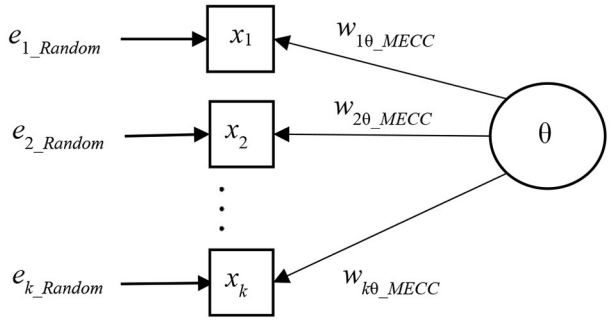

FIGURE 3 | MEC-corrected one-latent variable measurement model.

This conceptualization leads to short-cuts to estimate deflation-corrected reliability. These estimators are divided into two families as discussed above: on the one hand, *Rit* is replaced by a different coefficient in MECRs: on the other hand, an attenuation-corrected estimator of correlation is used in ACERs. These estimators are short-cuts in the sense that until a sound theoretical basis for a new way of thinking, defining, and estimating reliability is developed, these practical options would lead to a reasonable alternative to deflation-corrected estimates of reliability.

Theoretical Bases for the Deflation-Corrected Estimators of Reliability

The General (theoretical) bases for different families of DCERs discussed by Metsämuuronen (2022a,c,f) are based on alpha (eq. 3):

$$\rho_{\alpha\_wi\,\theta} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i w_{i\,\theta}\right)^2}\right), \tag{22}$$

theta (eq. 5):

$$\rho_{TH\_wi\,\theta} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k} w_{i\,\theta}^2}\right), \tag{23}$$

omega (eq. 6):

$$\rho_{\omega\_wi\,\theta} = \frac{\left(\sum_{i=1}^{k} w_{i\,\theta}\right)^2}{\left(\sum_{i=1}^{k} w_{i\,\theta}\right)^2 + \sum_{g=1}^{k}\left(1 - w_{i\,\theta}^2\right)}, \tag{24}$$

or rho (eq. 7):

$$\rho_{MAX\_wi\,\theta} = \frac{1}{1 + \frac{1}{\sum_{i=1}^{k}\left(w_{i\,\theta}^2/\left(1-w_{i\,\theta}^2\right)\right)}}, \tag{25}$$

where the notation $w_{i\,\theta}$ refers to the fact that the magnitude of the estimate depends on three things: characteristics of the weight factor (w), the item (i), and the score variable (θ) as a manifestation of the latent trait as discussed above. Other bases could also be used. However, using theta, omega, and rho outside of their traditional context is debatable. Here, it is assumed that the estimators *could* be used as independent estimators; this seems consistent with the general measurement model discussed above. Alternatively, we may think that the estimates we get using R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, or E_{AC} instead of the traditional λ_i are outcomes of renewed procedures on principal component and factor analysis where the factor loadings are, i.e., R_{PC} and G_2 instead of PMC (cl. ordinal theta by Zumbo et al., 2007).

The practical characteristics of the estimators are studied in the empirical section. From a theoretical viewpoint, in hypothetic extreme datasets with deterministic item discrimination in *all* items leading to $R_{PCi} = R_{PCj} \approx G_i = G_j = G_{2i} = G_{2j} = R_{ACi} = R_{ACj} = E_{ACi} = E_{ACj} \equiv 1$,[2] estimators based on rho (eq. 25) could not be used, because this would require division by zero, which is not defined. However, DCERs based on theta and omega (eqs. 23 and 24) would lead to perfect reliability ($REL = 1$):

$$\begin{aligned}\rho_{TH\_RPCi\,\theta}^{Max} &\approx \rho_{TH\_Gi\,\theta}^{Max} = \rho_{TH\_RACi\,\theta}^{Max} \\ &= k/(k-1)\ \left(1 - 1/k\right) \equiv 1\end{aligned} \tag{26}$$

and

$$\rho_{\omega\_RPCi\,\theta}^{Max} \approx \rho_{\omega\_Gi\,\theta}^{Max} = \rho_{\omega\_RACi\,\theta}^{Max} = (k)^2/\left((k)^2 + 0\right) \equiv 1. \tag{27}$$

The maximum value by the estimators based on alpha (eq. 22) is:

$$\begin{aligned}\rho_{\alpha\_RPCi\,\theta}^{Max} &\approx \rho_{\alpha\_Gi\,\theta}^{Max} = \rho_{\alpha\_RACi\,\theta}^{Max} \\ &= \frac{k}{k-1}\left(1 - \sum_{i=1}^{k}\sigma_i^2/\left(\sum_{i=1}^{k}\sigma_i\right)^2\right).\end{aligned} \tag{28}$$

Hence, estimators based on alpha can reach the value $\rho_{\alpha\_RPCi\,\theta}^{Max} \approx \rho_{\alpha\_Gi\,\theta}^{Max} = \rho_{\alpha\_RACi\,\theta}^{Max} = 1$ only when all item variances are equal ($\sigma_i = \sigma_j = \sigma$), that is, for instance, when the items are standardized. In the case

$$\begin{aligned}\rho_{\alpha\_RPCi\,\theta} &\approx \rho_{\alpha\_Gi\,\theta} = \rho_{\alpha\_Di\,\theta} \\ &= k/(k-1)\ \left(1 - k\sigma^2/(k\sigma)^2\right) \\ &= k/(k-1)\ \left(1 - 1/k\right) \equiv 1\end{aligned} \tag{29}$$

[2]Notably, *RPC* cannot reach a perfect 1. With enhanced procedures of the estimation by adding a very small number like 10^{-50} to each element of the logarithm and when the embedded PMC ≈ 1 such as 0.99999999, $RPC \approx 1$.

Notably, in the theoretical case, all the item–score correlations are equal to 0, and except for those based on omega, none of the estimators are defined. This is inherited from the original estimators: those that are not defined when all correlations or loadings are 0.

METHODOLOGY

Measurement Model and Estimators Used in the Empirical Section

In the empirical section, the characteristics of different types of DCERs are compared by varying the characteristics of w and i in a real-life setting with finite or small sample sizes. The general measurement model discussed above is applied in the empirical section. Formulae (22) to (25) are used as bases for the estimators. The raw score (θ_X) is used as the manifestation of θ and R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, and E_{AC} as weight factors. The estimators of correlation and their estimation are described in Appendix 1 in **Supplementary Material** (refer to details in, e.g., Metsämuuronen, 2022b). The estimates by the traditional estimators ρ_α, ρ_{TH}, ρ_ω, and ρ_{MAX} (eqs. 3, 6, 8, and 9), with their traditional original score variables (θ_X for alpha, θ_{PC} for theta, and θ_{FA} with ML estimation for omega and rho) and weight factor (*Rit* for alpha and λ_i for theta, omega, and rho), are used as benchmarks to the DCERs. With only two items with a wide-scale, principal axis factoring (PAF), instead of ML, is conducted to estimate the factor loadings.

In the empirical section, the estimators are named based on eqs. (22) to (25). For example, ρ_{MAX_RPCiX} refers to eq. (25) where the base is the formula of rho (ρ_{MAX}), the weight factor is R_{PC}, and the score variable is the raw score (θ_X). In the Figures and Tables, this is expressed as "RhoRPC." Similarly, the traditional estimators are referred to as "AlphaRit," "ThetaPC," "OmegaML," and "RhoML" or by an attribute "traditional" such as "Alpha traditional."

The estimators and estimates are also compared from the viewpoint of their capability of reflecting the population value. A simple statistic for this is used: the difference between the sample estimate and the population value (d). When $d > 0$, the true correlation is overestimated, and when $d < 0$, the sample estimate underestimates the population estimate. In the Figures and Tables, this difference related to a specific estimator is referred to as "dRhoRPC" and "dRho traditional".

Datasets Used and Tests Conducted in the Study

A real-world dataset of 4,022 nationally represented test-takers of a mathematics test with 30 binary items (FINEEC, 2018) is used as the "population". In the original dataset, $\rho_\alpha = 0.885$, $\rho_{TH} = 0.89$, $\rho_\omega = 0.887$, and $\rho_{MAX} = 0.895$; the difficulty levels of the items ranged $0.24 < p < 0.95$, with the average $\bar{p} = 0.63$; and item discrimination ranged $0.332 < Rit < 0.627$ with the average $\overline{Rit} = 0.481$.

Ten random samples with $n = 25$, 50, 100, and 200 test-takers were picked from the original dataset. These finite samples imitate different sizes of real-world sample sizes, ranging from a test for a large student group ($n = 200$) to classroom testing ($n = 25$). In each of the 10×4 datasets, 36 tests were produced by varying the number and difficulty levels of the items and the length of the scale of the score [$df(X)$ = number of categories in the scale−1] and the item [$df(g)$ = number of categories in the scale−1]. The polytomous items were constructed as sums of the original binary items. Thus, the datasets[3] consists of 14,880 partly related test items from 1,440 partly related tests with a varying number of test items ($k = 2$–30, $\bar{k} = 10.33$, SD 8.621) and test-takers ($n = 25$, 50, 100, and 200), number of categories in the items [$df(g) = 1$–14, $\overline{df(g)} = 4.57$, SD 3.480], and in the score [$df(X) = 10$–27, $\overline{df(X)} = 18.06$, SD 3.908], the average difficulty levels ($\bar{p}$= 0.50–0.76, $\bar{\bar{p}}$= 0.66. SD 0.052), and the lower bound of reliabilities ($\rho_\alpha = 0.55$–0.93, $\bar{\rho}_\alpha = 0.850$, SD 0.049).

RESULTS

Because previous studies related to DCERs have been fragmented, this study intends to offer a more systematic comparison of the estimators with a larger number of estimates. In doing so, five characteristics of DCERs are studied: their general tendencies in comparison with traditional estimators, their capability to reflect the population value, the effect of the sample size in the estimators, the effect of the number of categories in the score, and the effect of test difficulty. In what follows, mainly DCERs based on the form of omega ("deflation-corrected omega") are presented in the text, and all estimators in the comparison are collected in Appendix 2 in **Supplementary Material**.

General Tendencies of DCERs

Of the general tendencies of DCERs, three are highlighted. First, in comparison with the traditional estimators based on *Rit* and λ_i, all DCERs in the simulation give, in general, higher estimates. This is specifically true with binary datasets where all DCERs give systematically and consistently almost the same estimate, which is 0.07–0.09 units higher than the traditional estimates (**Table 1**; **Figure 4**; refer also to Appendix 2 in **Supplementary Material**). With binary items, all DCERs, irrespective of the base, suggest that the reliability of the (original) test would rather be 0.91–0.94 and not 0.85–0.88 as suggested by the traditional estimators. This higher magnitude of the estimates is caused by the less-deflated estimates of correlation with items of extreme difficulty level by the alternative estimators in comparison with PMC. Although the true reliability of the original real-life dataset is unknown, the unified voice of DCERs speaks of the possibility that they reflect the *same* (latent) true reliability. Notably, the differences between traditional estimates and those by DCERs are remarkably smaller than the ones in examples described by Gadermann et al. (2012) and

[3]The dataset of individual items ($n = 14{,}880$) including several indicators of item–score association is available in CSV format at http://dx.doi.org/10.13140/RG.2.2.10530.76482 and in SPSS format at http://dx.doi.org/10.13140/RG.2.2.17594.72641. The dataset of reliabilities ($n = 1{,}440$) is available in CSV format at http://dx.doi.org/10.13140/RG.2.2.30493.03040 and in SPSS format at http://dx.doi.org/10.13140/RG.2.2.27971.94241.

TABLE 1 | Average estimates of reliability and deviance from the population value in simulation.

	Traditionalestimators				MCER (R_{PC})				MCER (R_{REG})			
Base	**Alpha**	**Theta**	**Omega**	**Rho**	**Alpha**	**Theta**	**Omega**	**Rho**	**Alpha**	**Theta**	**Omega**	**Rho**
Estimate[a]	0.850	0.858	0.854	0.875	0.891	0.896	0.925	0.935	0.885	0.890	0.920	0.928
Deviation[b]	−0,016	−0,001	−0,012	0,012	−0,009	−0,002	−0,005	0,008	−0,005	0,001	−0,001	0,007
N	1,440	1,440	1,394	1,384	1,440	1,440	1,440	1,418	1,440	1,440	1,440	1,421
	MCER (G)				**MCER (D)**				**MCER (G_2)**			
Base	**Alpha**	**Theta**	**Omega**	**Rho**	**Alpha**	**Theta**	**Omega**	**Rho**	**Alpha**	**Theta**	**Omega**	**Rho**
Estimate[a]	0.831	0.834	0.893	0.904	0.789	0.796	0.873	0.883	0.905	0.910	0.933	0.942
Deviation[b]	−0,009	−0,005	−0,005	0,009	−0,010	−0,002	−0,005	0,009	−0,009	−0,001	−0,005	0,009
N	1,440	1,440	1,440	1,418	1,440	1,440	1,440	1,426	1,440	1,440	1,440	1,418
	MCER (D_2)				**ACER (R_{AC})**				**ACER (E_{AC})**			
Base	**Alpha**	**Theta**	**Omega**	**Rho**	**Alpha**	**Theta**	**Omega**	**Rho**	**Alpha**	**Theta**	**Omega**	**Rho**
Estimate[a]	0.884	0.890	0.920	0.930	0.891	0.897	0.924	0.934	0.901	0.906	0.930	0.939
Deviation[b]	−0,010	−0,002	−0,005	0,009	−0,007	0,001	−0,003	0,010	−0,006	0,001	−0,002	0,010
N	1,440	1,440	1,440	1,426	1,440	1,440	1,440	1,418	1,440	1,440	1,440	1,418

[a]*Average estimate.*
[b]*Average deviation between the sample and population estimates.*

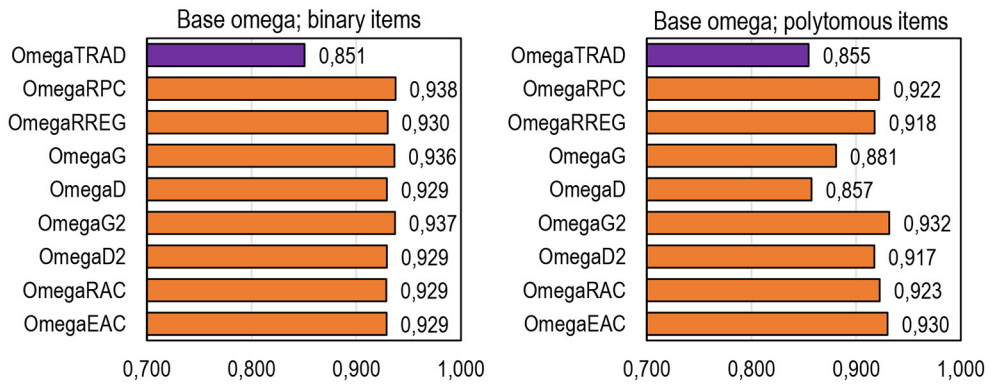

FIGURE 4 | Average estimates by DCERs based on the form of omega.

Metsämuuronen (2022a,c), and in extreme cases, the difference is reported to be 0.4–0.6 units of reliability. The smaller difference is caused by the fact that the datasets used in the simulation do not include extremely easy or extremely difficult items or tests.

Second, when the number of categories in the items exceeds 4, *G* and *D* tend to give an obvious underestimation of the item–score association (refer to, e.g., Metsämuuronen, 2021a). Hence, we obtain notably low estimates of reliability using alpha and theta as bases for the DCERs with items that have a wide scale (refer to **Figure 4**; Appendix 2 in **Supplementary Material**). In these cases, using the dimension-corrected estimators G_2 and D_2 would be better, with binary items $G = G_2$ and $D = D_2$. Using G_2 and D_2 as the linking factor with polytomous items seems to give largely the same magnitude of reliability as given by R_{PC} and R_{REG}.

Third, using rho as the base may lead to missing estimates, specifically with small sample sizes. Datasets with the smallest sample size in the simulation produce a remarkable number of deterministic patterns (6% of the estimates with $n = 25$) where the estimates based on rho are not defined. Then, factually, the number of estimates is 1,418 (instead of 1,440) for estimators based on rho (refer to **Table 1**). Small sample sizes are prone to produce not only deterministic patterns where rho cannot be calculated at all but also near-deterministic patterns leading to (artificially) high estimates. This characteristic seems to be inherited also to DCERs based on rho: the estimates based on rho with binary items (0.94–0.96) are suspiciously high in

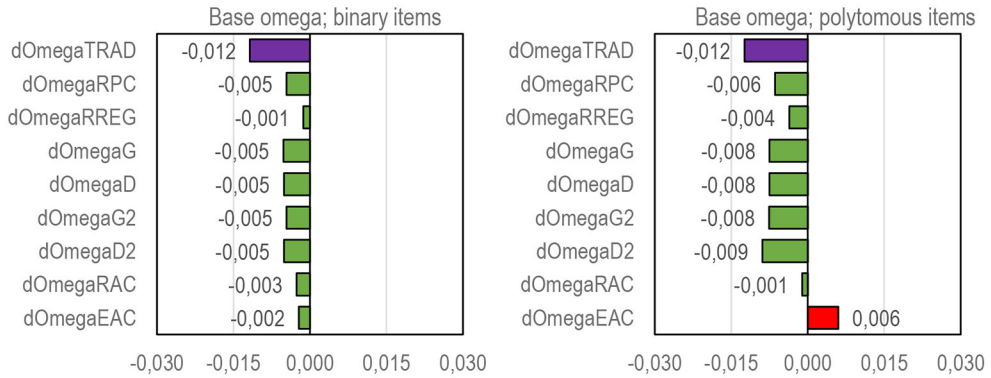

FIGURE 5 | Deviance between sample and population estimates by DCERs based on the form of omega.

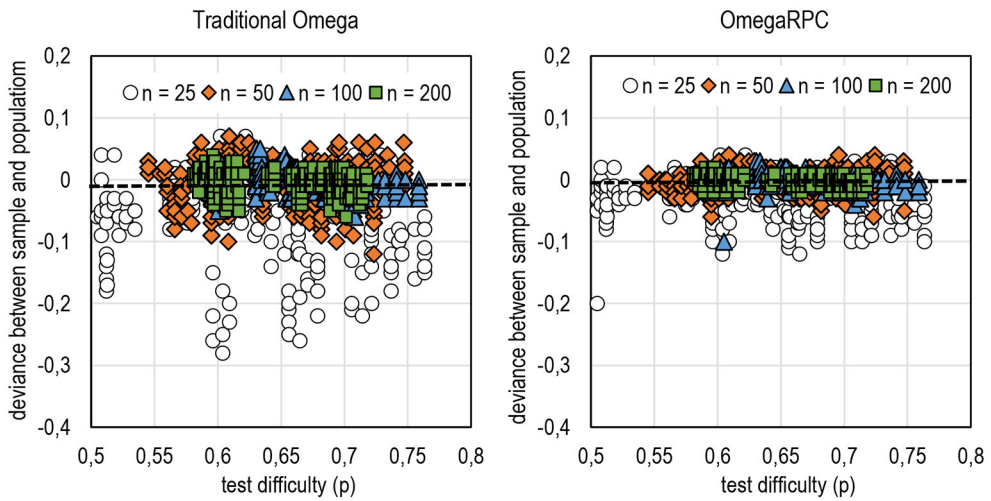

FIGURE 6 | Deviance between sample and population estimates by a DCER based on omega.

comparison with the estimators based on theta and omega (0.93–0.94; refer to Appendix 2 in **Supplementary Material**). This is related to the note by Aquirre-Urreta et al. (2019) that traditional rho tends to give overestimates with finite samples.

The Capability of DCERs to Reflect Population Reliability

Another aspect of the general tendencies is how well sample estimates reflect population estimates. This is illustrated in **Figure 5** and Appendix 2 in **Supplementary Material**, and four points are highlighted here. First, DCERs based on alpha, theta, and omega are conservative: they tend to produce estimates where the magnitude is lower than population reliability. In contrast, DCERs based on rho tend to be liberal: the estimates tend to overestimate population reliability, especially with binary items (refer to Appendix 2 in **Supplementary Material**). Second, sample estimators using E_{AC} as a linking factor tend to overestimate population reliability based on E_{AC}. Notably, the factual estimates of reliability seem not to be overestimated when E_{AC} is used (refer to **Figure 4** above). Third, estimators based on the form of theta and rho tend to be more stable than those using alpha and omega, theta in binary settings, and rho with polytomous settings (except when R_{AC} or E_{AC} are used as the linking factor; refer to Appendix 2 in **Supplementary Material**). In estimators based on theta and rho, the deviance between the sample and population estimates is generally around 0.001–0.002 units of reliability. With estimators based on alpha and omega, the deviance is around 0.01–0.02 units of reliability.

Fourth, although the general tendencies show only mild deviance between sample and population, single estimates in the sample may be far off the population value. **Figure 6** illustrates how widely the estimates may deviate from the population values, specifically with small sample sizes. The reason for the wide deviance with small sample sizes, specifically when using the traditional omega, is that even one test-taker may have a notable effect on changing the correlations between the item and score and, in some cases, even from positive (in the population) to

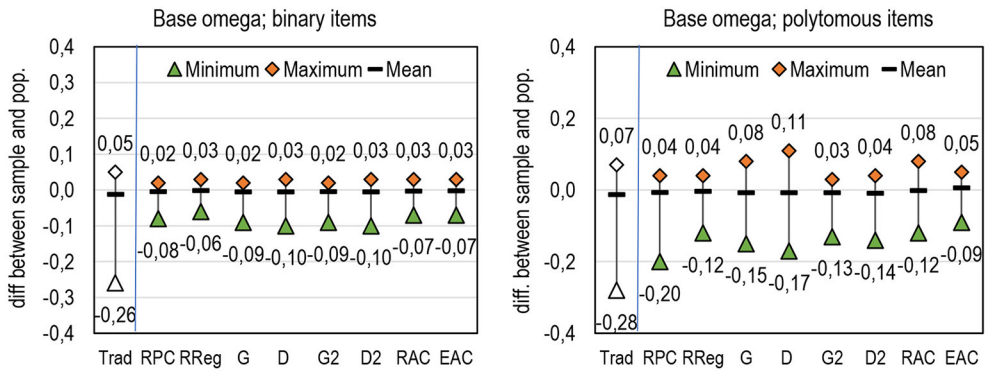

FIGURE 7 | Deviance between sample and population estimates in DCERs based on omega.

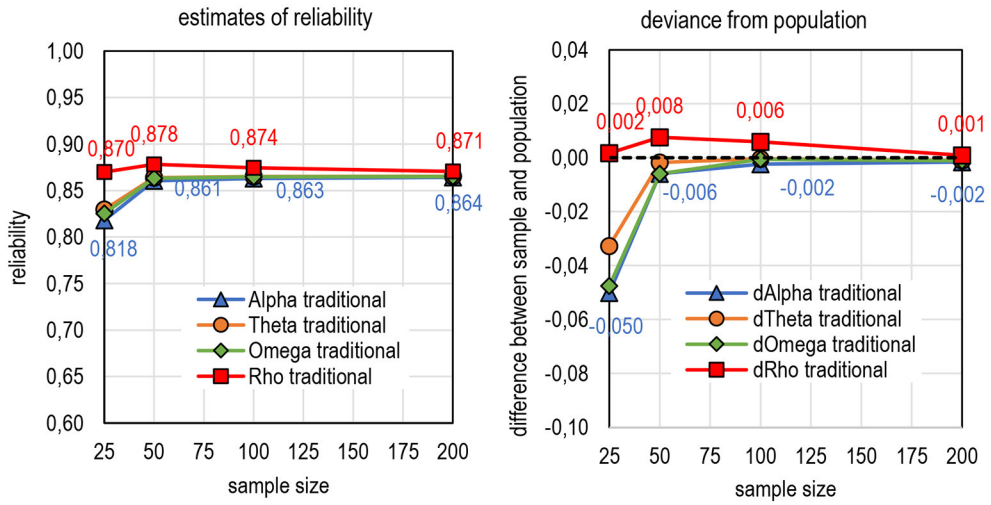

FIGURE 8 | Average estimates of reliability and deviance from the population by sample size.

negative in the sample (refer to examples in Metsämuuronen, 2022b).

Generally, except with estimators based on alpha, the deviance between the sample and population estimates seems notably smaller by DCERs than by traditional estimators (refer to **Figure 7**; Appendix 2 in **Supplementary Material**). Specifically, this is true with binary items. The traditional theta seems to give relatively more stable estimates even without correction for deflation. Notably, the wide range in deviance between the sample and population estimates with polytomous items when *G* or *D* are used as the linking factor and alpha as the base is caused by the fact that *G* and *D* tend to give obvious underestimation when the number of categories in item exceeds 3–4.

Effect of Sample Size on DCERs

As a benchmark to DCERs in **Figure 9**, **Figure 8** illustrates the behavior of the traditional estimators by sample size (refer to details in Appendix 2 in **Supplementary Material**). All the conservative estimators (alpha, theta, and omega) tend to give estimates that deviate notably from the population value when the sample size is very small ($n = 25$). When the sample size reaches $n = 50$, the estimates are relatively stable. Theta seems to be the most stable when it comes to reflecting the population value. The estimates by rho are higher than others, but it also tends to overestimate mildly population reliability (up to 0.008 units of reliability) with small sample sizes.

The estimates by DCERs differ notably depending on whether binary or polytomous items are used. With binary items, all DCERs give largely the same estimates, while with polytomous items, DCERs using *G* and *D* as the linking factor underestimate reliability irrespective of the sample size (refer to **Figure 9** and more details in Appendix 2 in **Supplementary Material**). In both cases, the estimates are stable when the sample size is $n = 50$ or higher. All the estimators underestimate population reliability with a very small sample size ($n = 25$).

It seems that DCERs give a notable advantage when the sample size is small. This is true specifically with binary items; the estimates by DCERs tend to be closer to the population value in

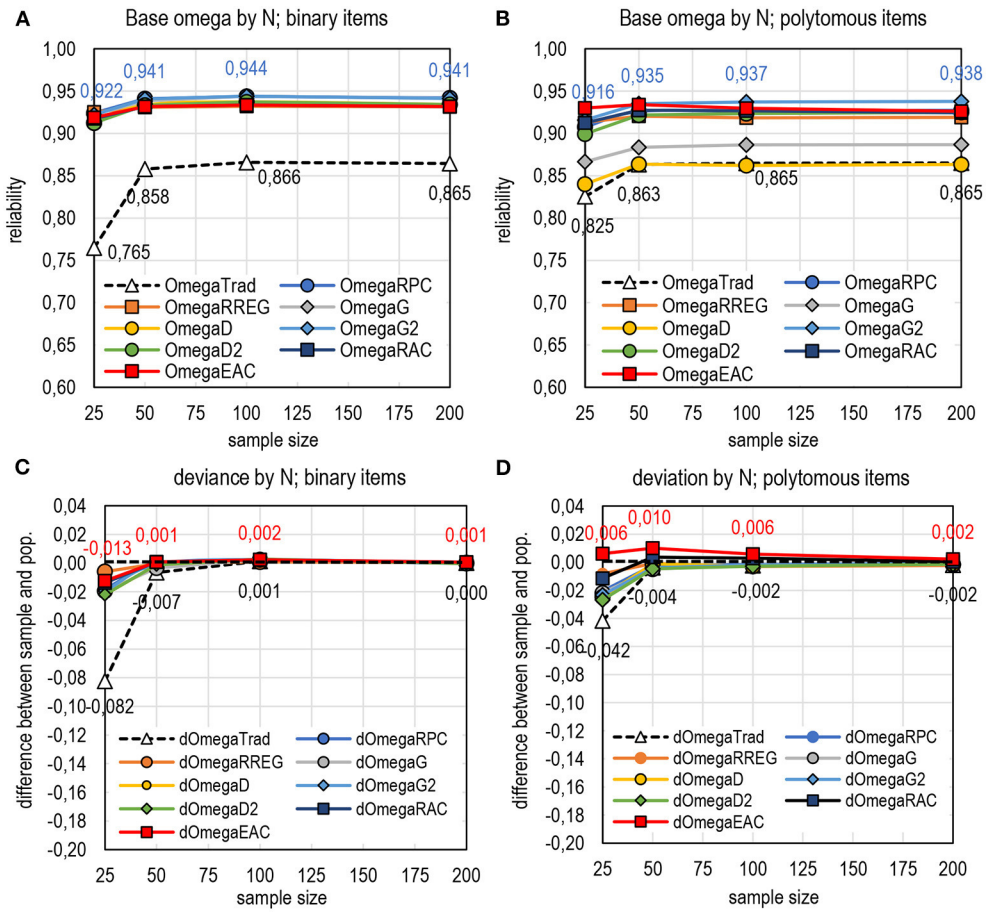

FIGURE 9 | The behavior of DCERs based on omega by sample size. **(A)** Base omega by N; binary items. **(B)** Base omega by N; polytomous items. **(C)** Deviance by N; binary items. **(D)** Deviation by N; polytomous items.

comparison with the traditional estimators. Omega would benefit the most by changing the linking factor. With polytomous items, DCERs using E_{AC} as the linking factor tend to overestimate the population value, although the factual estimates do not exceed the magnitude of the estimates using G_2 as the linking factor.

Traditional alpha, omega, and rho seem to benefit if the linking factor is changed from PMC to any of the item–score correlations used for comparison. The estimators using bi- and polyreg correlation coefficient (R_{REG}) with very small sample sizes seem to give more stable estimates than other estimators of correlation, and the estimates based on theta seem to be relatively stable even with small sample sizes and without changing the linking factor.

Effect of Number of Categories in the Score on DCERs

The dataset used in simulation is limited when it comes to the number of categories in the score variable. Because of the limitations in the original dataset, only scores with a number of categories ranging from 11 to 31 [$df(X) = 10$–30] could be used. However, it seems that all the estimators give stable estimates when the number of categories in the score exceeds 20 (**Figures 10a,b**).

Among the traditional estimators, alpha and omega seem quite unstable when the scale of the score is narrow [$df(X) < 15$], and the reliability of the population may be underestimated by more than 0.1 units (**Figure 10b**). From this viewpoint, the estimates by theta are notably closer to the population values as the reliability is underestimated by less than 0.06 units with binary items. The estimates by rho tends to overestimate reliability by up to 0.03 units with scores with a narrow scale, although the estimates tend to be rather stable with polytomous items even when the score has a narrow scale.

When it comes to DCERs, in general, those using a conservative base (alpha, theta, and omega) tend to underestimate population reliability less than the traditional estimators, specifically with scores with a narrow scale [$df(X) < 15$] and binary items, whereas those based on a liberal base (rho), tend to less overestimate population reliability than traditional estimators with short tests (**Figure 10b**; Appendix 2 in **Supplementary Material**). Although the DCERs that use

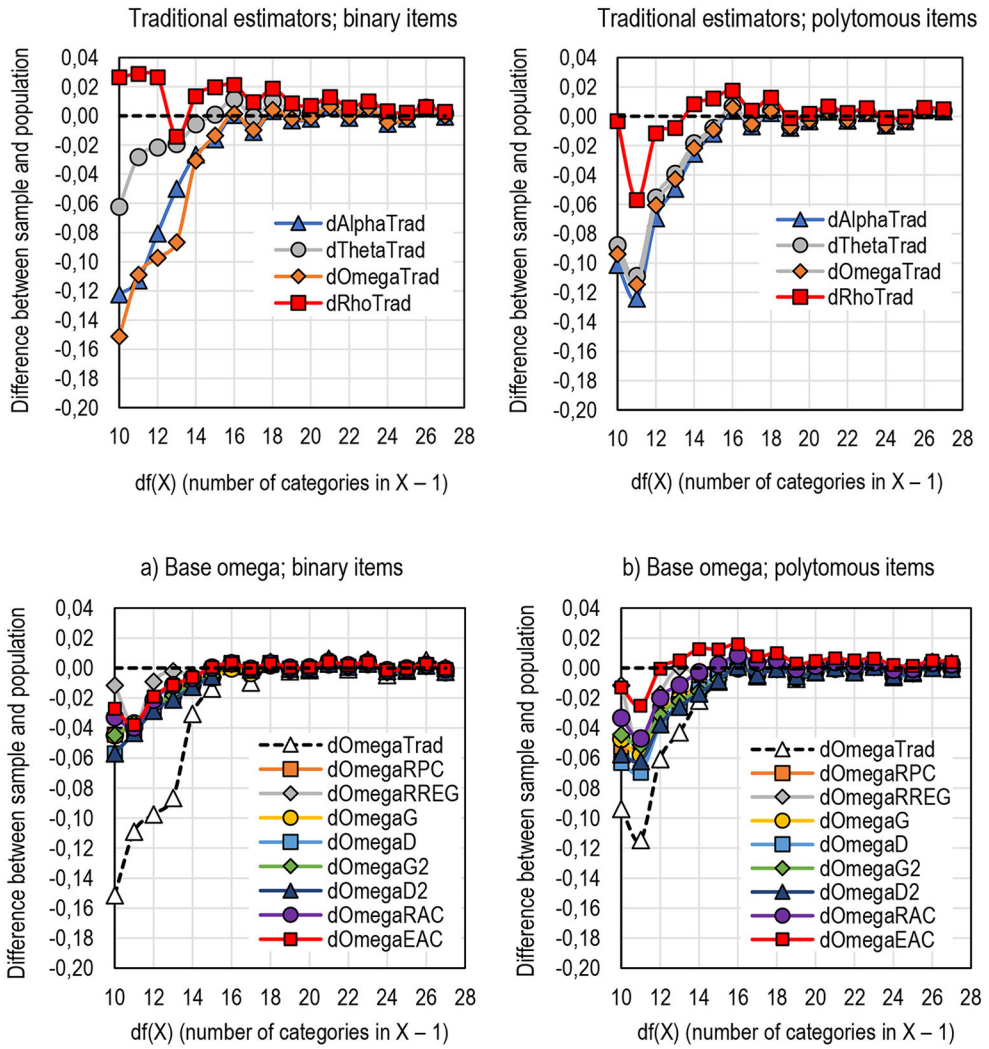

FIGURE 10 | The behavior of traditional estimators of reliability by the width of the score [df(X)]. The behavior of DCERs by the width of the score [df(X)]. **(a)** Base omega; binary items. **(b)** Base omega; polytomous items.

E_{AC} as the linking factor tend to overestimate reliability with polytomous items (refer to above), the estimates tend to be closest to the population value with polytomous items and very short tests [df(X) < 14].

Effect of Test Difficulty on DCERs

Lastly, the estimators are compared by their behavior for tests with different difficulty levels. Notably, the dataset used in the simulation does not allow comparing them with extremely difficult or extremely easy tests; in such tests, *Rit* is the most vulnerable. Still, some comparisons are conducted although the number of "difficult" (average proportion of correct answers in the items is $\bar{p}< 0.55$) and "easy" tests ($\bar{p}> 0.75$) is small. **Figures 11a,b** (refer also to Appendix 2 in **Supplementary Material**) illustrate the behavior of omega and the related DCERs regarding test difficulty, and three points are highlighted.

First, of the traditional estimators, alpha and omega tend to be more affected by test difficulty than theta and rho. Alpha and omega tend to underestimate reliability in both extremes. Theta seems relatively stable with binary items but is affected by test difficulty with polytomous items. Rho is stable, although it seems to overestimate reliability irrespective of test difficulty if the difficulty level is not extreme.

Second, with binary items, the magnitude of the estimates by DCERs tends to be notably higher and more stable than by the traditional estimators irrespective of test difficulty. A specific advantage of DCERs is with a test of extreme difficulty level where the traditional estimators tend to give lower values. This is specifically true with estimators based on alpha and omega; it seems that the traditional alpha and omega would benefit most by changing the linking factor.

Third, with polytomous items, using R_{AC} or E_{AC} as a linking factor seems to produce the most stable estimates irrespective

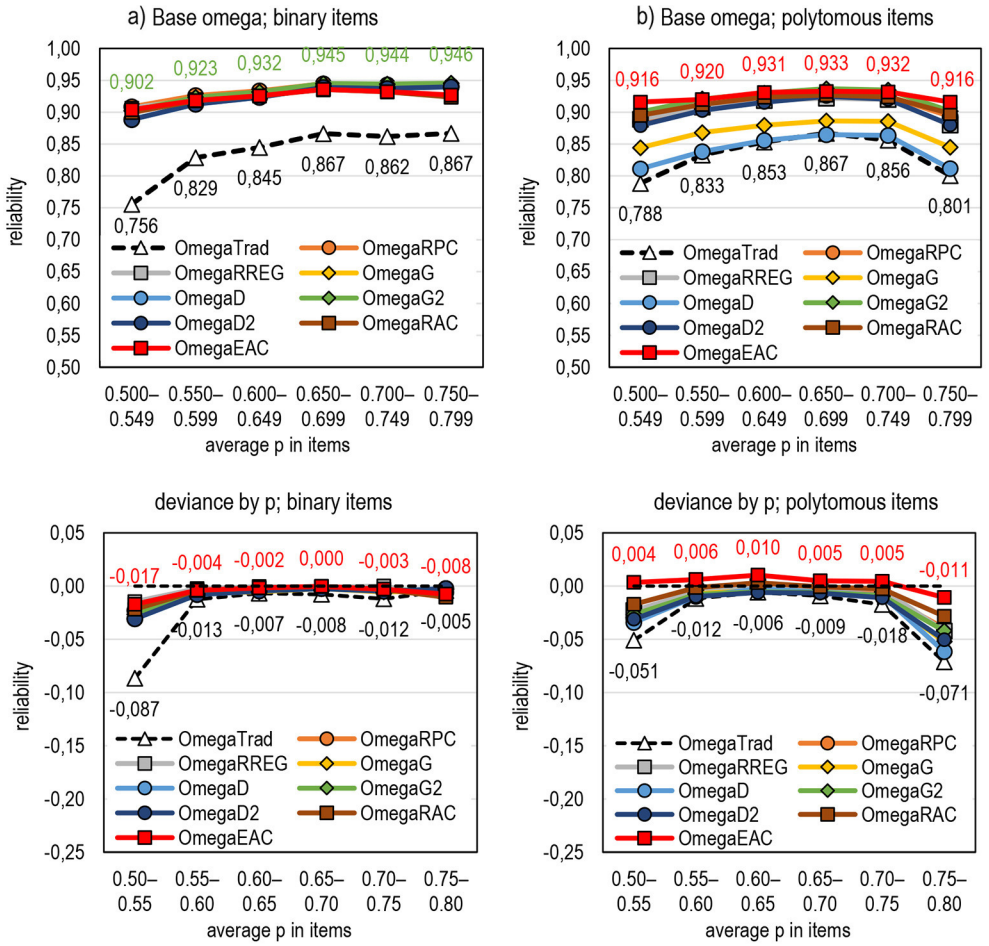

FIGURE 11 | The behavior of DCERs by test difficulty (the highest and traditional estimates are highlighted). **(a)** Base omega; binary items. **(b)** Base omega; polytomous items.

of the base used and test difficulty. E_{AC} tends to overestimate reliability mildly, but the factual estimates tend not to differ from those where G_2 is used. Except for the estimators that use D and G, the differences between the estimates are small.

CONCLUSIONS, DISCUSSION, AND RESTRICTIONS

Results in a Nutshell

The starting point of this article was two-fold. First, the empirical findings indicate that the estimates by the traditional estimators of reliability such as alpha, theta, omega, and rho tend to be deflated, and the magnitude of deflation may be remarkable with certain types of datasets, typically with tests including items of extreme difficulty level. Second, the main reason for the deflation in the estimates of reliability is the mechanical error related to estimates of the item–score correlation embedded in the widely used traditional estimators of reliability. The behavior of alternative estimators for *Rit* has been studied, and short-cut estimators of reliability that produce deflation-corrected estimates have been proposed based on replacing *Rit* with an alternative, which gives a radically smaller magnitude of deflation. Some of these alternatives are R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, and E_{AC}, which are discussed in the empirical section.

Different families of DCERs can be classified by the estimator used as the base, by score variables, and by weighting factors between item and score variable. Studies concerning DCERs have been either at a very initial stage, they have offered just some examples of the new possibility, they have been based on small datasets and have been fragmentary, or the simulations have made only a limited comparison of the behavior of some DCERs with their traditional counterparts. The aim of this study was to conduct a more systematic comparison of the behavior of different combinations of these elements and to typologize estimators that would show which estimator suits which situations. The simulation used here was based on finite sample sizes relevant to many real-life testing settings ($n \leq 200$). Although the simulation conducted and the dataset used have their restrictions, which will be discussed later, seven main outcomes may be presented here:

1) Regardless of the base and linking factor used, DCERs tend to give higher estimates than traditional estimators. This is because of higher magnitudes of the item–score correlations obtained by the alternative estimators than by the traditional *Rit*.
2) Not only are their estimates higher, DCERs seems to tend to produce estimates that are closer to the population value than the traditional estimators do.
3) Although the true reliability of the original real-life dataset is unknown, the unified voice of the DCERs, specifically with binary items, speaks that they reflect the same (latent) true reliability.
4) A specific advantage of DCERs seems to come from small sample size, short tests, and test with extreme difficulty levels and binary items. In these settings, the traditional conservative estimators (alpha, theta, and omega) may radically underestimate population reliability.
5) With binary items, all DCERs in the comparison seem to give almost an identical outcome that is notably higher than that given by the traditional estimators. The differences between DCERs are clearer with polytomous items.
6) Of the individual DCERs, those using G and D as the linking factor tend to be conservative with polytomous items, specifically if alpha and theta are used as the base. This is caused by the known characteristic of G and D to underestimate the item–score association in an obvious manner when the number of categories in the scale in an item exceeds 3–4. In these cases, instead of G and D, DCERs using dimension-corrected G and D (G_2 and D_2) as the linking factor give estimates with a magnitude close to the estimates by other estimators. Estimators using D_2 as the linking factor tend to give more conservative outcomes than G_2.
7) DCERs using E_{AC} as the linking factor offer a puzzle: although the magnitudes of the sample estimates are not higher than those given by the other DCERs, they tend to overestimate the *population* estimates using E_{AC} as the linking factor. This is specifically true when rho is used as the base with polytomous items. This uniquely reflects the relationship between the sample and population E_{AC}. A large population rarely leads to deterministic or near-deterministic patterns between two variables; small samples are more prone to these patterns, and the magnitude of the estimates by E_{AC} in a sample tends to be higher than in the population.

The characteristics of different combinations of the base and the linking factor are discussed in the section that follows.

Typology of Selected Deflation-Corrected Estimators of Reliability

Tables 2a,b summarize the typological characteristics of different combinations of the bases (alpha, theta, omega, and rho) and the weight factors (R_{PC}, R_{REG}, G, D, G_2, D_2, R_{AC}, and E_{AC}). Notably, all score variables discussed in the article (θ_X, θ_{PC}, θ_{FA}, θ_{IRT}, or θ_{NL}) are not covered in this study; the raw score (θ_X) was used in the simulation (of a comparison of other score variables; refer to Metsämuuronen, 2022a). The characteristics of the weight factors are studied elsewhere (e.g., Metsämuuronen, 2020a,b, 2021a,b, 2022b,d).

When it comes to the base of DCERs, the estimators based on alpha, theta, and omega are conservative; they tend to produce estimates that are underestimates of population reliability with small sample sizes. Estimators based on rho tend to be liberal; they tend to produce estimates that are overestimates of population reliability with small sample sizes. Estimators based on theta seem surprisingly stable, more stable than those by alpha and omega. Estimators based on rho are specifically vulnerable to deterministic patterns. In these patterns, estimates by rho cannot be calculated because of the undefined division by zero. Also, the estimates by rho are unstable with a near-deterministic pattern even in one item. These patterns are expected with small sample sizes. Hence, DCERs based on rho may not be suggested to be used with small sample sizes.

When it comes to weighting factors, R_{PC} and R_{REG} reflect a correlation between unobservable, theoretical constructions. Hence, DECRs using these coefficients as linking factors may lead to a kind of *theoretical* reliability that is not related to the factual score variable (refer to the critique by Chalmers, 2017). From this viewpoint, estimators based on G and D lead to more practical interpretations of reliability. That is, because G and, specifically, D strictly indicate the proportion of logically ordered test-takers in a test item after they are ordered by the score (refer to Metsämuuronen, 2021b), the DCERs using G or D reflect the proportion of logically ordered test-takers in all test items as a whole. For example, if the average D of all item–score correlations in a specific dataset is 0.7, it means that 85% of the test takers, that is, $p = 0.5 \times 0.70 + 0.5 = 0.85$ (refer to Metsämuuronen, 2021b), are logically ordered in all items as a whole after they are ordered by the score. Because of their conservative nature with polytomous items having more than three categories, DCERs based on G and D are suggested for tests with binary items and with polytomous items having less than four categories. The dimension-corrected versions of G and D (G_2 and D_2) can be used for binary and polytomous items and in a binary case, $G = G2$ and $D = D_2$.

Of the DCERs using attenuation-corrected estimators of correlation (R_{AC} and E_{AC}) as the linking factor, those using R_{AC} are more conservative than those using E_{AC}. This follows strictly from the behavior of R_{AC} and E_{AC}: except for the binary case where R_{AC} and E_{AC} give identical estimates, the estimates by E_{AC} tend to be higher than those by R_{AC} (refer to, e.g., Metsämuuronen, 2022d). Both seem to be somewhat liberal with small sample sizes especially with polytomous items, although the factual estimates do not seem to differ notably from the estimates by other DCERs. With binary items, ACERs tend to produce largely the same estimates as MCERs.

Based on the simulation, some initial recommendations concerning the usability of the DCERs may be summarized as follows; obviously, more specified simulations are needed, and these are discussed in the next section.

1) With small sample sizes ($n < 200$), using estimators based on rho is not recommendable; all DCERs based on rho as well

TABLE 2a | Typology of selected deflation-corrected estimators of reliability and their characteristics.

			RPC	RREG	G & D	G2 & D2
		General characteristics	• Reflects latent reliability, not strictly related to the observed score nor observed items • Leads to theoretical interpretation of reliability • Based on covariance • Suitable for binary and polytomous items • Not simple to calculate	• Reflects reliability of the observed score but uses non-observed items • Leads to partly theoretical interpretation of reliability • Based on regression model • Suitable for binary and polytomous items • Not simple to calculate	• Reflects reliability of observed score • Leads to practical interpretation of reliability • Based on probability • D more conservative than G • Suitable for binary items and polytomous items with < 3 categories • Simple to calculate manually	• Reflects reliability of the observed score but uses non-observed items • Leads to practical interpretation of reliability • Based on probability • Liberal nature; D_2 more conservative than G_2 • Suitable for binary and polytomous items • Simple to calculate manually
Base	Alpha	• Always underestimates population reliability • Very conservative in nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (rel = 1) when wi = 1, and $\sigma_i = \sigma_j$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i RPC_{i\theta}\right)^2}\right)$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i RREG_{i\theta}\right)^2}\right)$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i G_{i\theta}\right)^2}\right)$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i G_{2i\theta}\right)^2}\right)$
	Theta	• Maximizes alpha • Conservative nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (rel = 1) when wi = 1	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}RPC_{i\theta}^2}\right)$	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}RREG_{i\theta}^2}\right)$	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}G_{i\theta}^2}\right)$	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k}G_{2i\theta}^2}\right)$
	Omega	• Always higher than alpha • Least conservative nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (rel = 1) when wi = 1	$\frac{\left(\sum_{i=1}^{k}RPC_{i\theta}\right)^2}{\left(\sum_{i=1}^{k}RPC_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-RPC_{i\theta}^2\right)}$	$\frac{\left(\sum_{i=1}^{k}RREG_{i\theta}\right)^2}{\left(\sum_{i=1}^{k}RREG_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-RREG_{i\theta}^2\right)}$	$\frac{\left(\sum_{i=1}^{k}G_{i\theta}\right)^2}{\left(\sum_{i=1}^{k}G_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-G_{i\theta}^2\right)}$	$\frac{\left(\sum_{i=1}^{k}G_{2i\theta}\right)^2}{\left(\sum_{i=1}^{k}G_{2i\theta}\right)^2+\sum_{g=1}^{k}\left(1-G_{2i\theta}^2\right)}$
	rho (maximal reliability)	• Maximizes omega • Liberal nature; may overestimate reliability with small sample sizes • Cannot be calculated if deterministic patterns even in one item • Cannot reach the perfect reliability (rel = 1) • Not the best option for small samples	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(RPC_{i\theta}^2/\left(1-RPC_{i\theta}^2\right)\right)}}$	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(RREG_{i\theta}^2/\left(1-RREG_{i\theta}^2\right)\right)}}$	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(G_{i\theta}^2/\left(1-G_{i\theta}^2\right)\right)}}$	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(G_{2i\theta}^2/\left(1-G_{2i\theta}^2\right)\right)}}$

MEC-corrected estimators.

TABLE 2b | Typology of selected deflation-corrected estimators of reliability and their characteristics; attenuation-corrected estimators.

			Attenuation-corrected estimators;Weight wi	
			RAC	**EAC**
		General characteristics	• Reflects reliability of the observed score but uses non-observed items • Leads to practical interpretation of reliability • Based on probability • May have a liberal nature • Tendency for slight overestimation with polytomous items • Safe to use with items with < 4 categories • Simple to calculate manually	• Reflects reliability of the observed score but uses non-observed items • Leads to practical interpretation of reliability • Based on probability • Very liberal nature • Tendency for overestimation with polytomous items • Safe to use with binary items • Simple to calculate manually
Base	Alpha	• Always underestimates population reliability • Very conservative in nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (REL = 1) when wi = 1, and $\sigma_i = \sigma_j$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i RAC_{i\theta}\right)^2}\right)$	$\frac{k}{k-1}\left(1-\frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i EAC_{i\theta}\right)^2}\right)$
	Theta	• Maximizes alpha • Conservative nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (REL = 1) when wi = 1	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k} RAC_{i\theta}^2}\right)$	$\frac{k}{k-1}\left(1-\frac{1}{\sum_{i=1}^{k} EAC_{i\theta}^2}\right)$
	Omega	• Always higher than alpha • Least conservative nature • Gives estimates even with small sample sizes • Reaches the perfect reliability (REL = 1) when wi = 1	$\frac{\left(\sum_{i=1}^{k} RAC_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} RAC_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-RAC_{i\theta}^2\right)}$	$\frac{\left(\sum_{i=1}^{k} EAC_{i\theta}\right)^2}{\left(\sum_{i=1}^{k} EAC_{i\theta}\right)^2+\sum_{g=1}^{k}\left(1-EAC_{i\theta}^2\right)}$
	Rho (maximal reliability)	• Maximizes omega • Liberal nature; may overestimate reliability with small sample sizes • Cannot be calculated if deterministic patterns even in one item • Cannot reach the perfect reliability (rel < 1) • Not the best option for small samples	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(RAC_{i\theta}^2/\left(1-RAC_{i\theta}^2\right)\right)}}$	$\frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(EAC_{i\theta}^2/\left(1-EAC_{i\theta}^2\right)\right)}}$

as the traditional estimators tend to give overestimates with small sample sizes.

2) With binary items, all DCERs based on the conservative estimators (alpha, theta, and omega) give more plausible estimates than the traditional estimators; the difference is in the interpretation of the linking factor. Using R_{PC} or R_{REG} leads to "theoretical reliability" as a benchmark for the traditional one and using G or D (and G_2 or D_2) leads to practical interpretation of the logical order of the test-takers; all these refer to the discrimination power of the score. Using R_{AC} or E_{AC} may give an interpretation closer to the original *Rit*, that is, attenuation-corrected alpha, theta, omega, or rho. Notably, with binary items, R_{AC} and E_{AC} produce identical outcomes.
3) With polytomous items, DCERs using G and D are not recommended to be used is the number of categories exceeds 3 (D) or 4 (G), or, if used, the estimates may be very conservative—the magnitude of the estimates may be even more deflated than of those by the traditional

alpha. Specifically, if the number of categories in the score is small but the sample size is large, D tends to be affected by the large number of tied cases and tends to underestimate the correlation, which is also reflected in the estimates of reliability. With polytomous items, using G_2 or D_2 seems to give estimates whose magnitude is closer to those by R_{PC} or R_{REG}. However, using G_2 and EAC may give a liberal estimate in comparison with R_{PC}, R_{REG}, D_2, and R_{AC}.

4) If alpha and theta are used, where the traditional item–score correlation is originally used as default, as the bases for DCERs, attenuation-corrected *Rit* (R_{AC}) could be a natural alternative for *Rit*. Then, the "attenuation corrected alpha" or "attenuation corrected theta" could be reported as a benchmark as a side of the traditional alpha or theta. Using E_{AC} could enhance the outcome by also allowing non-linearity in the association between items and score. Obviously, the other alternative estimators could also be used; then, we could report "MEC-corrected alpha" or "deflation-corrected alpha" as a benchmark.
5) If using omega and rho as the bases for DCERs, three options may be worth considering. First, a renewed process of producing factor loadings may be considered; for DCERs, the factor loadings should be some of the alternative estimators of item–score correlation instead of (essentially) *Rit*. Second, another option to estimate the reliability of the factor score variables would be to estimate just the factor score variable by traditional factor analysis to produce an "optimal linear combination" and to use alternative estimators of item–score correlation in the DCERs irrespective of factor loadings. Third, in line with the general approach used in the article, the formulae of omega and rho could be used in DCERs to estimate the reliability of various types of score variables irrespective of the factor analysis. Systematic studies on these options would be beneficial.

Practical Calculation of DCERs

To give a practical example of calculating the DCERs discussed in this article, a specific national-level dataset with exceptionally easy items (n = 7,770) discussed by Metsämuuronen (2022b; 2022f; 2022g; originally in Metsämuuronen and Ukkola, 2019) and referred to in sections "From prediction formulae to coefficient alpha" and "From alpha, theta, omega, and rho to deflation-corrected reliability" is used here as an example. Originally, the test was a screening test of proficiency in the language used in the factual test; only test-takers with second language status were expected to make mistakes in the test items. Descriptive statistics of the dataset are collected in **Table 3a**, principal component and factor loadings for the traditional theta, omega, and rho in **Table 3b**, estimates of item–score correlation by selected estimators of correlation in **Table 3c**, and derivatives of the correlations for the traditional and deflation-corrected coefficients of alpha in **Table 3d**. Estimates of reliability are collected in **Table 3e**.

TABLE 3a | Descriptive statistics of the test items from Metsämuuronen and Ukkola (2019) (N = 7,770).

Item (g)	Range	Mean	*p*	Std. deviation	Variance
g1	0–1	0.96	0.96	0.186	0.0348
g2	0–1	0.98	0.98	0.126	0.0160
g3	0–1	0.99	0.99	0.088	0.0078
g4	0–1	0.91	0.91	0.287	0.0824
g5	0–2	1.78	0.89	0.610	0.3715
g6	0–1	0.98	0.98	0.122	0.0150
g7	0–2	1.97	0.985	0.211	0.0446
g8	0–2	1.98	0.99	0.169	0.0285
SUM					0.6004
Score	3–11	10.57	0.961	0.875	0.7650

TABLE 3b | Principal component and factor loadings.

	Principal component loadings and derivatives		Factorloadings and derivatives			
Item	λ_{PC}	λ^2_{PC}	λ_{MLE}	λ^2_{MLE}	$1-\lambda^2_{MLE}$	$\lambda^2_{MLE}/(1-\lambda^2_{MLE})$
g1	0.447	0.200	0.276	0.076	0.924	0.082
g2	0.430	0.185	0.260	0.068	0.932	0.073
g3	0.605	0.366	0.471	0.222	0.778	0.285
g4	0.468	0.219	0.291	0.085	0.915	0.093
g5	0.204	0.042	0.111	0.012	0.988	0.012
g6	0.375	0.141	0.213	0.045	0.955	0.048
g7	0.288	0.083	0.160	0.026	0.974	0.026
g8	0.633	0.401	0.512	0.262	0.738	0.355
SUM		1.636	2.294		7.204	0.974

TABLE 3c | Estimators of correlation between the item and raw score.

item	Rit	RPC	RREG	D	G	D2	G2	RAC	EAC
g1	0.351	0.677	0.436	0.791	0.857	0.791	0.857	0.551	0.551
g2	0.268	0.618	0.375	0.779	0.846	0.779	0.846	0.489	0.489
g3	0.283	0.696	0.408	0.858	0.911	0.858	0.911	0.603	0.603
g4	0.458	0.736	0.529	0.789	0.834	0.789	0.834	0.603	0.603
g5	0.746	0.931	0.732	0.952	0.979	0.958	0.982	0.921	0.923
g6	0.260	0.602	0.364	0.766	0.831	0.766	0.831	0.477	0.477
g7	0.327	0.702	0.425	0.832	0.897	0.943	0.976	0.568	0.567
g8	0.373	0.760	0.457	0.877	0.924	0.961	0.983	0.680	0.693

For the traditional alpha, theta, omega, and rho, their original score variable is used: a raw score for alpha, a principal component (PC) score for theta, and an ML estimate (MLE) of the factor score for omega and rho. For DCERs, the raw score is used as the manifestation of the latent variable; Metsämuuronen (2022f) shows examples of using PC and factor scores in calculations.

TABLE 3d | Derivatives of the estimators of correlation between an item and a raw score.

Item	VAR(g)	Rit × s	RPC × s	D × s	G × s	D2 × s	G2 × s	RAC × s	EAC × s
g1	0.035	0.065	0.126	0.147	0.160	0.147	0.160	0.103	0.103
g2	0.016	0.034	0.078	0.098	0.107	0.098	0.107	0.062	0.062
g3	0.008	0.025	0.061	0.076	0.080	0.076	0.080	0.053	0.053
g4	0.082	0.131	0.211	0.226	0.239	0.226	0.239	0.173	0.173
g5	0.372	0.455	0.568	0.580	0.597	0.584	0.598	0.561	0.562
g6	0.015	0.032	0.074	0.094	0.102	0.094	0.102	0.058	0.058
g7	0.045	0.069	0.148	0.176	0.189	0.199	0.206	0.120	0.120
g8	0.028	0.063	0.128	0.148	0.156	0.162	0.166	0.115	0.117
SUM	0.600	0.874	1.395	1.546	1.630	1.587	1.658	1.245	1.248

TABLE 3e | Estimates of reliability.

	Traditionalestimator	DCERs with alternative weight factors and raw score (θ_X)							
Base	Traditionalweight (score)	RPC	RREG	D	G	D2	G2	RAC	EAC
Alfa	0.2450 (θ_X)	0.7901	0.4196	0.8556	0.8846	0.8703	0.8934	0.7004	0.7025
Theta	0.4444 (θ_{PC})	0.8686	0.5200	0.9368	0.9610	0.9494	0.9684	0.7779	0.7802
Omega	0.4221 (θ_{MLE})	0.8952	0.6925	0.9473	0.9669	0.9572	0.9729	0.8310	0.8323
Rho	0.4934 (θ_{MLE})	0.9287	0.7353	0.9605	0.9795	0.9757	0.9891	0.9012	0.9031

Using **Tables 3a,d** and eq. (2), the estimate of reliability by the traditional alpha is $\hat{\rho}_{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i\rho_{i\,\theta\,X}\right)^2}\right) = \frac{8}{7}\left(1 - \frac{0.6004}{0.874^2}\right) = 0.245$. Correspondingly, using **Table 3b** and eqs. (6), (8) and (9), the estimate by theta is $\hat{\rho}_{TH} = \frac{k}{k-1}\left(1 - \frac{1}{\sum_{i=1}^{k}\lambda_{i\,\theta\,PC}^2}\right) = \frac{8}{7}\left(1 - \frac{1}{1.636}\right) = 0.444$, the estimate by omega is $\hat{\rho}_{\omega} = \frac{\left(\sum_{i=1}^{k}\lambda_{i\,\theta\,MLE}\right)^2}{\left(\sum_{i=1}^{k}\lambda_{i\,\theta\,MLE}\right)^2 + \sum_{i=1}^{k}\left(1-\lambda_{i\,\theta\,MLE}^2\right)} = \frac{2.294^2}{2.294^2+7.204} = 0.422$, and the estimate by rho is $\hat{\rho}_{MAX} = \frac{1}{1+\frac{1}{\sum_{i=1}^{k}\left(\lambda_{i\,\theta\,MLE}^2/\left(1-\lambda_{i\,\theta\,MLE}^2\right)\right)}} = \frac{1}{1+\frac{1}{0.974}} = 0.493$.

Similarly, the estimates by DCERs can be calculated using eqs. (22) to (25) by applying different weight factors.[4] If RPC is used as the weight factor, deflation-corrected alpha, as an example, gives an estimate of $\hat{\rho}_{\alpha\_RPCi\,\theta\,X} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i RPC_{i\,\theta\,X}\right)^2}\right) = \frac{8}{7}\left(1 - \frac{0.6004}{1.395^2}\right) = 0.790$ and, if G is used as the linking factor, $\hat{\rho}_{\alpha\_Gi\,\theta\,X} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_i^2}{\left(\sum_{i=1}^{k}\sigma_i G_{i\,\theta\,X}\right)^2}\right) = \frac{8}{7}\left(1 - \frac{0.6004}{1.630^2}\right) = 0.885$. In both cases, the message is the same: the estimate by the traditional alpha is radically deflated; instead of 0.24, the level of reliability is most probably closer to 0.79–0.85. Deflation-corrected thetas vary, 0.778–0.968, deflation-corrected omegas vary, 0.831–0.973, and deflation-corrected rhos vary, 0.901–0.989. These are notably higher than the deflated traditional theta (0.444), omega (0.422), and rho (0.493). In these kinds of datasets with extreme difficulty levels, DCERs may give a notable advantage in estimating the true reliability.

Known Limitations and Suggestions for Further Studies

The paradigm of deflation-correction in the estimates of reliability is still in the early stage. We do not know yet much about the new types of estimators of reliability. The simulation conducted in this article has obvious limits: only small sample sizes were used, the latent reliability was not controlled as is a norm in Monte Carlo simulations, the score variables was restricted only to raw score, tests with more than 30 and less than 10 categories in the score were missing, and no tests with extreme difficulty level or very short tests were not included in the simulation. Further investigation of such settings would be beneficial. Also, by far, only limited estimators of correlations as alternatives for *Rit* have been studied.

[4]The derivatives of the coefficients of correlation for DCERs based on theta, omega, and rho are not seen in **Tables 3b–d**. These are, however, easy to calculate from the original correlations in **Table 3c**, in the same manner done in **Table 3b**. Estimates by RREG seem notably lower than the other estimates of correlation; in what follows, these are taken as underestimates.

One obvious need of the new paradigm is to create a sound theoretical base for DCERs. From this viewpoint, DCERs based on omega and rho may be easier to argue for: the theoretical base discussed in eqs. (16) to (21) may be used as a sufficient conceptual or theoretical basis for DCERs. However, many traditional estimators are strictly based on variances, observed variance and error variance, leading to use of the traditional item–score correlation, which leads to deflation. The alternative estimators discussed in this article are mainly short-cuts replacing *Rit* in the process. However, if we want to create or develop an estimator such as ρ_{BS}, ρ_{FR}, ρ_{KR20}, and ρ_{α} from scratch and to avoid embedding *Rit* in the formulae, would the estimator still look like in the traditional formulae?

Another obvious restriction of the study is that only estimators from the classical test theory were discussed. A relevant question is, how applicable the results would be with estimators of reliability within Generalizability Theory (G-Theory; chronologically, e.g., Cronbach et al., 1972; Shavelson et al., 1989; Shavelson and Webb, 1991; Brennan, 2001, 2010; Vispoel et al., 2018a,b; Clayson et al., 2021), confirmatory factor analysis (CFA) or structural equation modeling (SEM refer to, e.g., Raykov and Marcoulides, 2006; Green and Yang, 2009b), and IRT and Rasch modeling (refer to estimators in e.g., Verhelst et al., 1995; Holland and Hoskens, 2003; Kim and Feldt, 2010; Cheng et al., 2012; Kim, 2012; Milanzi et al., 2015)? Except for the estimators developed for CFA and SEM analysis, in all cases, the possible deflation in the estimates is not as obvious as with the classical estimators, because the latter can be expressed using Rit and principal and factor loadings that are obviously deflated. Estimators using factor loading (as is a tradition in the basic CFA and SEM) are most probably prone to severe deflation because factor loadings are prone to deflation.

In G-Theory, the challenge is that, first, *two* types of estimators are used: the generalizability coefficient and the dependability coefficient; the former is low when interindividual rankings are inconsistent, and the latter is low when measurements from same individuals are inconsistent (refer to condensed discussion in Clayson et al., 2021). Although the former is more comparable with classical estimators such as coefficient alpha, we do not know the possible *mechanics* of deflation in these estimators. Second, in estimating the reliability within the framework of G-Theory, variance components are radically more complicated than when using classical estimators (refer to Brennan, 2001; Vispoel et al., 2018a; Clayson et al., 2021). Furthermore, Vispoel et al. (2018a) noted that failing to consider each source of measurement variance can result in overestimation of reliability. Hence, systematic theoretical and empirical studies are needed to confirm the possible sources of deflation in estimates by G-Theory.

In Rasch and IRT modeling, the estimation of reliability is often based on such concepts as "person separation" in Rasch models (Andrich and Douglas, 1977; Andrich, 1982; Wright and Masters, 1982) or "information function" in wider IRT models (refer to, e.g., McDonald, 1999; Cheng et al., 2012; Milanzi et al., 2015). These are not necessarily prone to deflation in an obvious manner. However, what *is* known is that the estimator called Accuracy of Measurement (MAcc) discussed by Verhelst et al. (1995) with a one-parameter logistic model tends to be severely affected by the form of distribution of the score; when the score variable is notably skewed, that is, when the test is either extremely easy or difficult to the target population, the estimates may even be far off the range of reliability (refer to the empirical examples in Metsämuuronen, 2022g).[5] If we assume that the estimates may be deflated in the estimators of reliability within the IRT modeling, two possible sources would be worth studying: the formulae themselves may not be effective or the estimates for item discrimination (*a*-parameter) often needed in the estimation would be deflated. With MAcc, it seems obvious that the operationalization of error variance of the score should be reconsidered (refer to Metsämuuronen, 2022g). Systematic studies, in this regard, would be beneficial.

Using score variance as a basis of reliability within the classical test theory leads easily to item–score correlation, which leads to deflation. If we want to avoid using variances as the base for reliability, one option for reconceptualizing reliability discussed by Metsämuuronen (2022a) is to define "perfect reliability" (*REL* = 1) as a condition where the score can discriminate test-takers in all items in a deterministic manner in the spirit of Guttman's scalogram (Guttman, 1950). This is related to the estimators of reliability within the non-parametric IRT modeling (NIRT; Mokken, 1971) where the coefficient H by Loevinger (1948) indicates homogeneity in the dataset and deviance from the deterministic pattern or so-called "Guttman-homogeneity" (refer to Molenaar and Sijtsma, 1984). This could lead to (correctly) detecting perfect reliability by DCERs based on theta and omega using R_{PC}, G, G_2, R_{AC}, and E_{AC} as the linking factors (see eqs. 22–25). D could be used as the linking factor in defining restrictions in Monte Carlo simulations: 90% of logically ordered test-takers in all items, after they are ordered by the score, lead to omegaD $= 0.9^2 = 0.81$ and 80% to omegaD $= 0.8^2 = 0.64$. Other options could be based on "sufficiency of information" (Smith, 2005), "person separation" (Andrich and Douglas, 1977; Andrich, 1982; Wright and Masters, 1982; refer also to "Rasch reliability" in Linacre, 1997; Clauser and Linacre, 1999), the "information function" (refer to, e.g., McDonald, 1999; Cheng et al., 2012; Milanzi et al., 2015) discussed in item response theory (IRT) settings, or "person-fit" within the paradigm of NIRT (refer to, e.g., Meijer et al. (1995).

The final note for further studies comes from the fact that the extended family of DCERs also includes estimators such as the

[5]In the specific dataset of achievement in the instruction language of a test in mathematics ($n = 7{,}770$) with extremely easy items and radically non-normal distribution discussed by Metsämuuronen (2022a,c,f) and re-analyzed above, the estimate by MAcc (Verhelst et al., 1995, pp. 99–100) was obviously out of range (MAcc $= -5.89$), while the traditional alpha $= 0.245$, theta $= 0.444$, omega $= 0.422$, and rho $= 0.493$, although deflated, were within the range of reliability. In April 2022, this specific dataset was re-analyzed by the teams of Milanzi et al. (2015) and Cheng et al. (2012) using the estimators they suggested in their articles. The results will be reported later. In this case, it would also be informative to apply Foster's (2021) enhanced KR20 developed for non-normal datasets such as exponential distributions in the score.

ordinal alpha and ordinal theta proposed by Zumbo et al. (2007). Other less known estimators may also be included. Ordinal alpha and theta are based on changing the inter-item matrices of PMCs by matrices of $R_{PC}s$ instead of changing the linking factor itself. It is expected that the estimates by ordinal alpha and theta would be identical with those by the theta RPC and alpha RPC discussed in this article, because the estimates using the traditional formula of alpha and an alternative computational form using the matrices of inter-item correlations are identical. However, it is not known whether estimates by factor analysis using the matrix of RPCs would lead to factor loadings that are R_{PC}s. If the estimates are identical, it would be easy to obtain DCERs based on omega and rho using traditional procedures simply by changing the inter-item matrix of *Rit*s to the matrix of R_{PC}s, *G*s, or *D*s, for instance. However, if the loadings are still (essentially) *Rit*s, calculated using the mechanics of PMC, it could be valuable to develop new procedures for FA/PCA so that the factor loadings needed in DCERs would be, factually, R_{PC}s, *G*s, or *D*s, for instance, as discussed above.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://dx.doi.org/10.13140/RG.2.2.10530.76482 http://dx.doi.org/10.13140/RG.2.2.17594.72641. http://dx.doi.org/10.13140/RG.2.2.30493.03040 http://dx.doi.org/10.13140/RG.2.2.27971.94241.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.891959/full#supplementary-material

REFERENCES

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR20 index, and the Guttman scale response pattern. *Educ. Res. Perspect.* 9, 95–104.

Andrich, D., and Douglas, G. A. (1977). *Reliability: Distinctions Between Item Consistency and Subject Separation With the Simple Logistic Model.* Paper presented at the Annual Meeting of the American Educational Research Association, April 4–8, 1977, New York City.

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika* 18, 1–14. doi: 10.1007/BF02289023

Aquirre-Urreta, M., Rönkkö, M., and McIntosh, C. N. (2019). A Cautionary note on the finite sample behavior of maximal reliability. *Psychologic. Methods*. 24, 236–252. doi: 10.1037/met0000176

Armor, D. (1973). Theta reliability and factor scaling. *Sociologic. Methodol.* 5, 17–50. doi: 10.2307/270831

Bentler, P. M. (1968). Alpha-maximized factor analysis (Alphamax): its relation to alpha and canonical factor analysis. *Psychometrika* 33, 335–345. doi: 10.1007/BF02289328

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika* 74, 137–143. doi: 10.1007/s11336-008-9100-1

Bravais, A. (1844). Analyse Mathematique. Sur les probabilités des erreurs de situation d'un point. (Mathematicalanalysis. Of the probabilities of the point errors). *Mémoires présentés par divers savants à l'Académie Royale des Siences de l'Institut de France*. 9, 255–332.

Brennan, R. L. (2001). *Generalizability Theory: Statistics for Social Science and Public Policy*. New York, NY: Springer-Verlag.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Appl. Measure. Educ.* 24, 1–21. doi: 10.1080/08957347.2011.532417

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *Br. J. Psychol.* 3, 296–322. doi: 10.1111/j.2044-8295.1910.tb00207.x

Chalmers, R. P. (2017). On misconceptions and the limited usefulness of ordinal alpha. *Educ. Psychologic. Measure.* 78, 1056–1071. doi: 10.1177/0013164417727036

Chan, D. (2008). "So why ask me? are self-report data really that bad?," in *Statistical and Methodological Myths and Urban Legends*, eds C. E. Lance and R. J. Vanderberg (London: Routledge), pp. 309–326. https://doi.org/10.4324/9780203867266

Cheng, Y., Yuan, K.-H., and Liu, C. (2012). Comparison of reliability measures under factor analysis and item response theory. *Educ. Psychologic. Measure.* 72, 52–67. doi: 10.1177/0013164411407315

Cho, E., and Chun, S. (2018). Fixing a broken clock: a historical review of the originators of reliability coefficients including Cronbach's alpha. *Survey Res.* 19(2), 23–54.

Cho, E., and Kim, S. (2015). Cronbach's coefficient alpha: well known but poorly understood. *Organization. Res. Method.* 18, 207–230. doi: 10.1177/1094428114555994

Clauser, B., and Linacre, J. M. (1999). Relating cronbach and rasch reliabilities. *Rasch Measure. Transact.* 13, 696.

Clayson, P. E., Carbine, K. A., Baldwin, S. A., Olsen, J. A., and Larson, M. J. (2021). Using generalizability theory and the ERP Reliability Analysis (ERA) Toolbox for assessing test-retest reliability of ERP scores part 1: algorithms, framework, and implementation. *Int. J. Psychophysiol.* 166, 174–187. doi: 10.1016/j.ijpsycho.2021.01.006

Cleff, T. (2019). "Applied statistics and multivariate data analysis for business and economics," in *A modern approach using SPSS, Stata, and Excel.* New York, NY: Springer.

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications. *J. Appl. Psychol.* 78, 98–104. doi: 10.1037/0021-9010.78.1.98

Cramer, D., and Howitt, D. (2004). *The Sage Dictionary of Statistics: A Practical Resource for Students.* London: SAGE Publications, Inc.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnum, N. (1972). *The Dependability of Behavioral Measures: Theory of Generalizability for Scores and Profiles*. London: John Wiley.

Dunn, T. J., Baguley, T., and Brunsden, V. (2013). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *Br. J. Psychol.* 105, 399–412. doi: 10.1111/bjop.12046

Edwards, A. A., Joyner, K. J., and Schatschneider, C (2021). A simulation study on the performance of different reliability estimation methods. *Educ. Psychologic. Measure*. 81, 1089–1117. doi: 10.1177/0013164421994184

Falk, C. F., and Savalei, V. (2011). The relationship between unstandardized and standardized alpha, true reliability, and the underlying measurement model. *J. Personal. Assess.* 93, 445–453. doi: 10.1080/00223891.2011.594129

Feldt, L. S. (1975). Estimation of reliability of a test divided into two parts of unequal length. *Psychometrika* 40, 557–561. doi: 10.1007/BF02291556

Feldt, L. S., and Brennan, R. L. (1989). "Reliability," in *Educational Measurement American council of Education Series of Higher Education*, ed R. L. Linn. Phoenix: Oryx Press.

FINEEC (2018). *National assessment of learning outcomes in mathematics at grade 9 in 2002 (Unpublished dataset opened for the re-analysis 18.2.2018).* Finnish National Education Evaluation Centre (FINEEC).

Foster, R. C. (2021). KR20 and KR21 for some nondichotomous data (it's not just Cronbach's alpha). *Educ. Psychologic. Measure.* 81, 1172–1202. doi: 10.1177/0013164421992535

Gadermann, A. M., Guhn, M., and Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Pract. Assess. Res. Eval.* 17, 1–13. doi: 10.7275/n560-j767

Gilmer, J. S., and Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lengths. *Psychometrika* 48, 99–111. doi: 10.1007/BF02314679

Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Am. Statistic. Assoc.* 49, 732–764. doi: 10.1080/01621459.1954.10501231

Graham, J. M. (2006). Congeneric and (essentially) tau–equivalent estimates of score reliability: What they are and how to use them. *Educ. Psychologic. Measure.* 66, 930–944. doi: 10.1177/0013164406288165

Green, S. B., and Yang, Y. (2009a). Commentary on coefficient alpha: a cautionary tale. *Psychometrika* 74, 121–135. doi: 10.1007/s11336-008-9098-4

Green, S. B., and Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: an alternative to coefficient alpha. *Psychometrika* 74, 155–167. doi: 10.1007/s11336-008-9099-3

Greene, V. L., and Carmines, E. G. (1980). Assessing the reliability of linear composites. *Sociologic. Methodol.* 11,160–17. doi: 10.2307/270862

Gulliksen, H. (1950). *Theory of Mental Tests*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Guttman, L. (1941). "The qualifications of a class of attributes: a theory and method of scale construction," in *The Prediction of Personal Adjustment. Social Science Research Council,* ed P. Horst, pp. 321–345.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika* 10, 255–282. doi: 10.1007/BF02288892

Guttman, L. (1950). "The basis for scalogram analysis," in *Measurement and Prediction*, eds S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfield, S. A. Star, and J. A. Clausen (London: Princeton University Press).

Hancock, G. R., and Mueller, R. O. (2001). "Rethinking construct reliability within latent variable systems," in *Structural Equation Modeling: Present and Future — A Festschrift in honor of Karl Jöreskog*, eds R. Cudeck, S. du Toit, and D. Sörbom (New York, NY: Scientific Software International, Inc), p. 195–216.

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

Heise, D., and Bohrnstedt, G. (1970). Validity, invalidity, and reliability. *Sociologic. Methodol.* 2, 104–129. doi: 10.2307/270785

Henrysson, S. (1963). Correction of item–total correlations in item analysis. *Psychometrika* 28, 211–218. doi: 10.1007/BF02289618

Hoekstra, R., Vugteveen, J., Warrens, M. J., and Kruyen, P. M. (2019). An empirical analysis of alleged misunderstandings of coefficient alpha. *Int. J. Soc. Res. Methodol.* 22, 351–364. doi: 10.1080/13645579.2018.1547523

Holland, P. W., and Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika* 68, 123–149. doi: 10.1007/BF02296657

Horst, P. (1951). Estimating the total test reliability from parts of unequal length. *Educ. Psychologic. Measure.* 11, 368–371. doi: 10.1177/001316445101100306

Jackson, P. H., and Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: algebraic lower bounds. *Psychometrika* 42, 567–578. doi: 10.1007/BF02295979

Jackson, R. W. B., and Ferguson, G. A. (1941). *Studies on the Reliability of Tests*. Toronto: Department of Educational Research, University of Toronto.

Kaiser, H. F., and Caffrey, J. (1965). Alpha factor analysis. *Psychometrika* 30, 1–14. doi: 10.1007/BF02289743

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *J. Educ. Psychol.* 30, 17–24. doi: 10.1037/h0057123

Kendall, M. G. (1948). *Rank Correlation Methods (1st ed)*. London: Charles Griffin and Co Ltd.

Kendall, M. G. *(1938)*. A new measure of rank correlation. *Biometrika* 30, 81–93. doi: 10.2307/2332226

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika* 77, 153–162. doi: 10.1007/s11336-011-9238-0

Kim, S., and Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Educ. Rev.* 11, 179–188. doi: 10.1007/s12564-009-9062-8

Kuder, G. F., and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika* 2, 151–160. doi: 10.1007/BF0228 8391

Lavrakas, P. J. (2008). "Attenuation," in *Encyclopedia of Survey Methods*, ed P. J. Lavrakas (London: Sage Publications, Inc).

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika* 62, 245–249. doi: 10.1007/BF02295278

Li, H., Rosenthal, R., and Rubin, D. B. (1996). Reliability of measurement in psychology: from spearman-brown to maximal reliability. *Psychologic. Methods* 1, 98–107. doi: 10.1037/1082-989X.1.1.98

Linacre, J. M. (1997). KR-20 / Cronbach alpha or Rasch reliability: which tells the "truth"? *Rasch Measure. Transact.* 11, 580–581.

Livingston, S. A., and Dorans, N. J. (2004). *A graphical approach to item analysis.* (Research Report No. RR-04-10). Educational Testing Service.

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychologic. Bull.* 45, 507–529. doi: 10.1037/h0055827

Lord, F. M. (1958). Some relations between Guttman's principal component scale analysis and other psychometric theory. *Psychometrika* 23, 291–296. doi: 10.1002/j.2333-8504.1957.tb00073.x

Lord, F. M., Novick, M. R., and Birnbaum, A. (1968). *Statistical Theories of Mental Test Scores*. Boston, MA: Addison–Wesley Publishing Company.

McDonald, R. P. (1970). Theoretical canonical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *Br. J. Mathematic. Statistic. Psychol.* 23, 1–21. doi: 10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1985). *Factor Analysis and Related Methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

McDonald, R. P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychologic. Methods* 23, 412–433. doi: 10.1037/met0000144

Meijer, R. R., Sijtsma, K., and Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT Model. *Appl. Psychologic. Measure.* 19, 323–335. doi: 10.1177/014662169501900402

Metsämuuronen, J. (2016). Item–total correlation as the cause for the underestimation of the alpha estimate for the reliability of the scale. *GJRA—Glob. J. Res. Anal.* 5, 471–477.

Metsämuuronen, J. (2017). *Essentials of Research Methods in Human Sciences*. New Delhi: SAGE Publications, Inc.

Metsämuuronen, J. (2018). Algebraic reasons why item-rest correlation underestimates item discrimination power more than item-test correlation. *[Preprint].* doi: 10.13140/RG.2.2.25568.94728

Metsämuuronen, J. (2020a). Somers' D as an alternative for the item–test and item–rest correlation coefficients in the educational measurement settings. *Int. J. Educ. Methodol.* 6, 207–221. doi: 10.12973/ijem.6.1.207

Metsämuuronen, J. (2020b). Dimension-corrected Somers' D for the item analysis settings. *Int. J. Educ. Methodol.* 6, 297–317. doi: 10.12973/ijem.6.2.297

Metsämuuronen, J. (2020c). Generalized discrimination index. *Int. J. Educ. Methodol.* 6, 237–257. doi: 10.12973/ijem.6.2.237

Metsämuuronen, J. (2021a). Goodman–Kruskal gamma and dimension-corrected gamma in educational measurement settings. *Int. J. Educ. Methodol.* 7, 95–118. doi: 10.12973/ijem.7.1.95

Metsämuuronen, J. (2021b). Directional nature of Goodman-Kruskal gamma and some consequences. Identity of Goodman-Kruskal gamma and Somers delta, and their connection to Jonckheere-Terpstra test statistic. *Behaviormetrika* 48, 2. doi: 10.1007/s41237-021-00138-8

Metsämuuronen, J. (2022a). Deflation-corrected estimators of reliability. *Front. Psychol.* 12, 748672. doi: 10.3389/fpsyg.2021.748672

Metsämuuronen, J. (2022b). Effect of various simultaneous sources of mechanical error in the estimators of correlation causing deflation in reliability. Seeking the

best options of correlation for deflation-corrected reliability. *Behaviormetrika* 49, 91–130. doi: 10.1007/s41237-022-00158-y

Metsämuuronen, J. (2022c). Attenuation-corrected reliability and some other MEC-corrected estimators of reliability. *Appl. Psychologic. Measure.* (in printing)

Metsämuuronen, J. (2022d). Artificial systematic attenuation in eta squared and some related consequences. attenuation-corrected eta and eta squared, negative values of eta, and their relation to pearson correlation. *Behaviormetrika* 12, 62. doi: 10.1007/s41237-022-00162-2

Metsämuuronen, J. (2022e). Essentials of visual diagnosis of test items. Logical, illogical, and anomalous patterns in tests items to be detected. *Pract. Assess. Res. Eval.* 27, 5. doi: 10.7275/n0kf-ah40

Metsämuuronen, J. (2022f). How to obtain the most error-free estimate of reliability? Eight sources of underestimation of reliability. *Pract. Assess. Res. Eval.* 27, 10. doi: 10.7275/7nkb-j673

Metsämuuronen, J. (2022g). Reliability for a score compiled from multiple booklets with equated scores. *[Preprint].* http://dx.doi.org/10.13140/RG.2.2.20880.69120/1

Metsämuuronen, J., and Ukkola, A. (2019). Alkumittauksen menetelmällisiä ratkaisuja (Methodological solutions of zero level assessment). Publications 18:2019. Finnish Education Evaluation Centre. [in Finnish, abstract in English]. Available online at: https://karvi.fi/app/uploads/2019/08/KARVI_1819.pdf (accessed June 11, 2022).

Milanzi, E., Molenberghs, G., Alonso, A., Verbeke, G., and De Boeck, P. (2015). Reliability measures in item response theory: manifest vs. latent correlation functions. *Br. J. Mathematic. Statistic. Psychol.* 68, 43–64. doi: 10.1111/bmsp.12033

Mokken, R.J. (1971). *A Theory and Procedure of Scale Analysis*. Berlin: de Gruyter.

Molenaar, I. W., and Sijtsma, K. (1984). Internal consistency and reliability in Mokken's nonparametric item response model. *Tijdschrift voor Onderwijsresearch* 9, 257–268.

Moltner, A., and Revelle, W. (2015). *Find the Greatest Lower Bound to Reliability*. Available online at: http://personality-project.org/r/psych/help/glb.algebraic.html (accessed June 11, 2022).

Moses, T. (2017). "A review of developments and applications in item analysis," in *Advancing Human Assessment. The Methodological, Psychological and Policy Contributions of ETS*, eds R. Bennett and M. von Davier (New York, NY: Educational Testing Service. Springer Open), pp. 19–46.

Novick, M. R., and Lewis, C. (1967). Coefficient alpha and the reliability of composite measurement. *Psychometrika* 32, 1–13. doi: 10.1007/BF02289400

Olvera Astivia, O. L., Kroc, E., and Zumbo, B. D. (2020). The role of item distributions on reliability estimation: the case of Cronbach's coefficient alpha. *Educ. Psychologic. Measure.* 80, 825–846. doi: 10.1177/0013164420903770

Pearson, K (1896). VII. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophic. Transact. Royal Soc. London* 187, 253–318. doi: 10.1098/rsta.1896.0007

Pearson, K. (1900). I. Mathematical contributions to the theory of evolution. VII. On the correlation of characters not quantitatively measurable. *Philosophic. Transact. Royal Soc. A. Mathematic. Physic. Eng. Sci.* 195, 1–47. doi: 10.1098/rsta.1900.0022

Pearson, K. (1903). I. Mathematical contributions to the theory of evolution. — XI. On the influence of natural selection on the variability and correlation of organs. *Philosophic. Transact. Royal Soc. A. Mathematic. Physic. Eng. Sci.* 200, 1–66. doi: 10.1098/rsta.1903.0001

Pearson, K. (1905). *On the General Theory of Skew Correlation and Non-Linear Regression*. London: Dulau and Co. Available online at: https://onlinebooks.library.upenn.edu/webbin/book/lookupid?key=ha100479269 (accessed June 11, 2022).

Pearson, K. (1909). On a new method of determining correlation between a measured character A, and a character B, of which only the percentage of cases wherein B exceeds (or falls short of) a given intensity is recorded for each grade of A. *Biometrika* 7, 96–105. doi: 10.1093/biomet/7.1-2.96

Pearson, K. (1913). On the measurement of the influence of "broad categories" on correlation. *Biometrika* 9, 116–139. doi: 10.1093/biomet/9.1-2.116

Raju, N. S. (1977). A generalization of coefficient alpha. *Psychometrika* 42, 549–565. doi: 10.1007/BF02295978

Raykov, T. (1997a). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence for fixed congeneric components. *Multivariate Behav. Res.* 32, 329–354. doi: 10.1207/s15327906mbr3204_2

Raykov, T. (1997b). Estimation of composite reliability for congeneric measures. *Appl. Psychologic. Measure.* 21, 173–184. doi: 10.1177/01466216970212006

Raykov, T. (2004). Estimation of maximal reliability: a note on a covariance structure modeling approach. *Br. J. Mathematic. Statistic. Psychol.* 57, 21–27. doi: 10.1348/000711004849295

Raykov, T., and Marcoulides, G. A. (2017). Thanks coefficient alpha, we still need you! *Education. Psychologic. Measure.* 79, 200–210. doi: 10.1177/0013164417725127

Raykov, T., and Marcoulides**,** G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Struct. Eq. Model. Multidisciplin. J.* 13, 130–141. doi: 10.1207/s15328007sem1301_7

Raykov, T., and Marcoulides**,** G. A. (2010). *Introduction to Psychometric Theory*. London: Routledge.

Raykov, T., West, B. T., and Traynor, A. (2014). Evaluation of coefficient alpha for multiple componentmeasuring instruments in complex sample designs. *Struct. Eq. Model.* 22(3), 429–438. doi: 10.1080/10705511.2014.936081

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behav. Res.* 14, 57–74. doi: 10.1207/s15327906mbr1401_4

Revelle, W. (2015). *Alternative estimates of Test Reliability*. Available online at: http://personality-project.org/r/html/guttman.html (accessed June 11, 2022).

Revelle, W. (2021). *Classical Test Theory and the Measurement of Reliability*. Available online at: http://www.personality-project.org/r/book/Chapter7.pdf (accessed June 11, 2022).

Revelle, W., and Condon, D. M. (2018). *Reliability from α to ω: A tutorial. [Preprint].* doi: 10.31234/osf.io/2y3w9

Revelle, W., and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika* 74, 145–154. doi: 10.1007/s11336-008-9102-z

Rulon, P. J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educ. Rev.* 9, 99–103.

Sackett, P. R., Lievens, F., Berry, C., M., and Landers, R. N. (2007). A cautionary note on the effect of range restriction on predictor intercorrelations. *J. Appl. Psychol.* 92, 538–544. doi: 10.1037/0021-9010.92.2.538

Sackett, P. R., and Yang, H. (2000). Correction for range restriction: An expanded typology. *J. Appl. Psychol.* 85, 112–118. doi: 10.1037/0021-9010.85.1.112

Schmidt, F. L., and Hunter, J. E. (2015). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (3rd ed.). London: SAGE Publications.

Schmidt, F. L., Shaffer, J. A., and Oh, I.-S. (2008). Increased accuracy for range restriction corrections: implications for the role of personality and general mental ability in job and training performance. *Personnel Psychol.* 61, 827–868. doi: 10.1111/j.1744-6570.2008.00132.x

Shavelson, R. J., and Webb, N. M. (1991). *Generalizability Theory: A Primer*. London: SAGE Publications, Inc.

Shavelson, R. J., Webb, N. M., and Rowley, G. L. (1989). Generalizability theory. *Am. Psychol.* 44, 922–932. doi: 10.1037/0003-066X.44.6.922

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

Smith, J. K. (2005). Reconsidering reliability in classroom assessment and grading. *Educ. Measure. Issues Pract.* 22, 26–33. doi: 10.1111/j.1745-3992.2003.tb00141.x

Somers, R. H. (1962). A new asymmetric measure of correlation for ordinal variables. *Am. Sociologic. Rev.* 27, 799–811. doi: 10.2307/2090408

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1422689

Spearman, C. (1910). Correlation computed with faulty data. *Br. J. Psychol.* 3, 271–295. doi: 10.1111/j.2044-8295.1910.tb00206.x

Stouffer, S. A. (1950). *Measurement and Prediction. Studies in Social Psychology in World war II, Vol IV*. Princeton, NJ: Princeton university press.

Ten Berge, J. M. F., and Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika* 43, 575–579. doi: 10.1007/BF02293815

Thompson, G. H. (1940). Weighting for battery reliability and prediction. *Br. J. Mathematic. Statistic. Psychol.* 30/4, 357–360. doi: 10.1111/j.2044-8295.1940.tb00968.x

Trizano-Hermosilla, I., and Alvarado, J. M. (2016). Best alternatives to Cronbach's alpha reliability in realistic conditions: congeneric and asymmetrical measurements. *Front. Psychol.* 7, 769. doi: 10.3389/fpsyg.2016.00769

Verhelst, N. D., Glas, C. A. W., and Verstralen, H. H. F. M. (1995). *One Parametric Logistic Model OPLM.* Arnhem, NL: CITO. doi: 10.1007/978-1-4612-4230-7_12

Vispoel, W. P., Morris, C. A., and Kilinc, M. (2018a). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychologic. Method.* 23, 1–26. doi: 10.1037/met0000107

Vispoel, W. P., Morris, C. A., and Kilinc, M. (2018b). Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *J. Personal. Assess.* 100, 53–67. doi: 10.1080/00223891.2017.1296455

Warrens, M. J. (2015). Some relationships between Cronbach's alpha and the Spearman-Brown formula. *J. Classific.* 32, 127–137. doi: 10.1007/s00357-015-9168-0

Warrens, M. J. (2016). A comparison of reliability coefficients for psychometric tests that consist of two parts. *Adv. Data Anal. Classific.* 10, 71–84. doi: 10.1007/s11634-015-0198-6

Woodhouse, B., and Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: a search procedure to locate the greatest lower bound. *Psychometrika* 42, 579–591. doi: 10.1007/BF02295980

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement*. San Diego: Mesa Press.

Xiao, L., and Hau,. K.-T. (2022). Performance of coefficient alpha and its alternatives: effects of different types of non-normality. *Educ. Psychologic. Measure.* 22, 240. doi: 10.1177/00131644221088240

Yang, H. (2010). "Factor loadings," in *Encyclopedia of Research Design*, ed N. J. Salkind (London: SAGE Publications), pp. 480–483.

Yang, Y., and Green, S. B. (2011). Coefficient alpha: a reliability coefficient for the 21st Century? *J. Psychoeducat. Assess.* 29, 377–392. doi: 10.1177/0734282911406668

Zinbarg, R. E., Revelle, W., Yovel, I., and Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each *o*ther and two alternative conceptualizations of reliability. *Psychometrika* 70, 123–133. doi: 10.1007/s11336-003-0974-7

Zumbo, B. D., Gadermann, A. M., and Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *J. Mod. Appl. Statistic. Methods* 6, 21–29. doi: 10.22237/jmasm/1177992180

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 | Frontiers in Psychology

TYPE Original Research
PUBLISHED 14 October 2022
DOI 10.3389/fpsyg.2022.899165

OPEN ACCESS

EDITED BY
Begoa Espejo,
University of Valencia, Spain

REVIEWED BY
Joshua Ray Tanzer,
Lifespan, United States
Katerina M. Marcoulides,
University of Minnesota Twin Cities,
United States

*CORRESPONDENCE
Caroline Keck
caroline.keck@UGent.be

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

RECEIVED 18 March 2022
ACCEPTED 31 August 2022
PUBLISHED 14 October 2022

CITATION
Keck C, Mayer A and Rosseel Y (2022) Overview and evaluation of various frequentist test statistics using constrained statistical inference in the context of linear regression. *Front. Psychol.* 13:899165. doi: 10.3389/fpsyg.2022.899165

Overview and evaluation of various frequentist test statistics using constrained statistical inference in the context of linear regression

Caroline Keck[1]*, Axel Mayer[2] and Yves Rosseel[1]

[1]Department of Data Analysis, Ghent University, Ghent, Belgium, [2]Psychological Methods and Evaluation, Bielefeld University, Bielefeld, Germany

Within the framework of constrained statistical inference, we can test informative hypotheses, in which, for example, regression coefficients are constrained to have a certain direction or be in a specific order. A large amount of frequentist informative test statistics exist that each come with different versions, strengths and weaknesses. This paper gives an overview about these statistics, including the Wald, the LRT, the Score, the $\bar{F}$- and the D-statistic. Simulation studies are presented that clarify their performance in terms of type I and type II error rates under different conditions. Based on the results, it is recommended to use the Wald and $\bar{F}$-test rather than the LRT and Score test as the former need less computing time. Furthermore, it is favorable to use the degrees of freedom corrected rather than the naive mean squared error when calculating the test statistics as well as using the $\bar{F}$- rather than the $\bar{\chi}^2$-distribution when calculating the p-values.

KEYWORDS

informative hypothesis testing, constrained statistical inference, informative test statistics, type I error rates, naive mean squared error, corrected mean squared error, $\bar{F}$-distribution, $\bar{\chi}^2$-distribution

Introduction

Imagine a researcher wants to examine a novel psychotherapy program. A randomized experiment is set up with three treatment groups. One is a control group ($X = 0$), one participates in an established, standard psychotherapy program ($X = 1$) and one participates in the novel psychotherapy program ($X = 2$). No covariates are considered. The researcher is interested in the group means of the dependent variable Y, which denotes the score on a mental health questionnaire. Studies like this are usually conducted to show the superiority of the novel treatment over the standard treatment, as well as the superiority of the standard treatment over the control group. Thus, the researcher assumes that $\mu_2 > \mu_1 > \mu_0$. However, following classical null hypothesis testing procedures, we usually first test a hypothesis like $H_0 : \mu_2 = \mu_1 = \mu_0$ against H_a: not H_0, that is, not all three means are equal.

If we can reject H_0 in favor of H_a, a second step often follows, in which we execute pairwise comparisons to determine which means are equal and which means are not equal. This implies multiple testing, which brings along the risk of an inflated type I error rate. The framework of constrained statistical inference (Silvapulle and Sen, 2005; Hoijtink, 2012) allows us to test so-called informative hypotheses, meaning that we can test the null hypothesis $H_0 : \mu_2 = \mu_1 = \mu_0$ against the ordered hypothesis $H_a : \mu_2 > \mu_1 > \mu_0$ in a single step. Thus, in contrast to classical null hypothesis testing, researchers have the advantages that they can formulate their hypotheses of interest directly, instead of making a detour via another hypothesis, while additionally avoiding to increase the risk for inflated type I error rates.

Informative hypothesis testing can be conducted by means of the Bayesian (see, e.g., Hoijtink et al., 2008; Hoijtink, 2012) as well as the frequentist (see, e.g., Barlow et al., 1972; Robertson et al., 1988; Silvapulle and Sen, 2005) approach, where the latter is the focus of this paper. The Bayesian approach is implemented in the R (R Core Team, 2020) package bain (Gu et al., 2020). The frequentist approach is implemented in SAS/STAT® by means of the PLM procedure (for instructions, see Chapter 87 of SAS Institute Inc., 2015) as well as in several R packages including restriktor (Vanbrabant, 2020) and ic.infer (Grömping, 2010). Recent work of Keck et al. (2021) also demonstrated how to integrate informative hypothesis testing into the EffectLiteR (Mayer and Dietzfelbinger, 2019) package.

Restriktor and ic.infer use a broad range of test statistics, which are presented in Silvapulle and Sen (2005). However, research in the field of constrained statistical inference often uses the famous $\bar{F}$-statistic (see, e.g., Kuiper and Hoijtink, 2010; Vanbrabant et al., 2015) and neglects the distance statistic (D-statistic). Furthermore, each test statistic comes in various versions, for example depending on which estimate is used for the mean squared error or the variance-covariance matrix, and oftentimes, it is not obvious which software program uses which test statistic. There are also different options regarding the distributions that can be used to compute the p-values ($\bar{\chi}^2, \bar{F}$). At the same time, small sample properties of informative test statistics are mostly unknown. Finally, simulation studies that examine the performance of informative test statistics are lacking in the constrained statistical inference literature.

The aim of this paper is twofold. First, we want to give an overview of a broad range of different informative test statistics, including the Wald test, the likelihood-ratio test (LRT), the Score test, the $\bar{F}$- and the D-statistic as well as their different versions. Second, we want to clarify how those test statistics perform when sample and effect sizes, hypotheses and the distribution used for calculating the p-values vary. Note that we only consider the regression setting, where all variables are observed. The paper is structured as follows: We start by presenting the univariate linear regression model to explain all necessary terminology that is used in the following section, where we define the test statistics. These test statistics include "regular" as well as informative test statistics to illustrate the link between them. We also discuss different versions of these test statistics. Subsequently, we report about simulation studies that we conducted. We introduce the design of the studies, that included a broad range of sample sizes as well as effect sizes, and we outline type I and type II error rates. We conclude with a short discussion. Supplementary materials are provided and will be referenced throughout the paper.

Univariate linear regression model

The univariate linear regression model for an observation i can be defined as:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \quad (1)$$

where y_i is the value of the response variable for observation $i = 1, 2, ..., n$, x_{i0} is 1 and $x_{i1}, ..., x_{ip}$ are the values of the p regressors for observation i, which are assumed to be fixed (in terms of repeated sampling). $\beta_0, ..., \beta_p$ are the regression coefficients and ε_i is a residual for observation i. In matrix notation, the model can be written as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is called the design matrix.

This regression model relies on several assumptions. First, we assume that the expected value of ε_i is zero. That is, $E(\varepsilon_i) = 0$ for all i. In matrix notation, this is expressed as $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, which implies that $E(\boldsymbol{y}) = \boldsymbol{X\beta}$, meaning that there is a linear relationship between $E(\boldsymbol{y})$ and the columns of $\boldsymbol{X}$. Second, we assume that x_i is non-stochastic and $\boldsymbol{X}$ is of full column rank. Third, we assume that the error term has a constant variance: $Var(\varepsilon_i) = \sigma_\varepsilon^2$ for all i. This implies that $Var(y_i) = \sigma_\varepsilon^2$ for all i. Fourth, we assume that the covariance of any two error terms is zero, that is $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all (i, j), where $i \neq j$.

The model can be estimated by means of different approaches such as ordinary least squares (OLS) or maximum likelihood (ML). It can be shown that under the presented assumptions, the OLS estimates of $\boldsymbol{\beta}$ are BLUE (best linear unbiased estimators, see, e.g., Seber and Lee, 2012). Using an example including four predictors, the following model is fitted:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, \quad (2)$$

and $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, and $\hat{\beta}_4$ are obtained via OLS estimation. We may be interested in hypotheses concerning a single parameter like $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ or we might be interested in hypotheses about nested model comparisons like $H_0 : \beta_1 = 0 \wedge \beta_2 = 0$ vs. $H_a : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \beta_4 \neq 0$. We can compute various important quantities that are used in hypothesis testing and that are characterized by a hat on top of it. Note that the hat indicates that estimation of the model parameters takes place in an unrestricted way, which will change once we test informative hypotheses. First, an unbiased

estimator for the mean squared error is:

$$\hat{\sigma}_{\varepsilon}^2 = \hat{S}^2_{corrected} = \frac{\widehat{RSS}}{n-k}, \tag{3}$$

where k is the column rank of $\boldsymbol{X}$ and $\widehat{RSS}$ is the estimated residual sum of squares $\sum_{i=1}^{n} \hat{e}_i^2$, where $\hat{e}_i = y_i - \hat{y}_i$ and $\hat{y}_i$ are the model predicted values of the response variable. Note that by considering k, we yield a small-sample correction for the mean squared error, as opposed to simply using:

$$\hat{S}^2_{naive} = \frac{\widehat{RSS}}{n}, \tag{4}$$

which corresponds to the maximum likelihood estimator of σ_{ε}^2.

The variance-covariance matrix of the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ is usually computed as:

$$VCOV(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\hat{\boldsymbol{I}}_1^{-1}, \tag{5}$$

where $\hat{\boldsymbol{I}}_1$ is the unit information matrix:

$$\hat{\boldsymbol{I}}_1 = \frac{1}{n\,\hat{S}^2_{corrected}}\boldsymbol{X}'\boldsymbol{X}. \tag{6}$$

Note that if certain model assumptions are violated, for example if the error term does not have a constant variance, robust versions of the standard errors (Huber, 1967; White, 1980) and the variance-covariance matrix (Zeileis, 2006) can be used.

We can also test hypotheses about linear or non-linear combinations of regression parameters, like $H_0 : \beta_1 + \beta_2 = 0 \wedge \beta_3 + \beta_4 = 0$ vs. $H_a : \beta_1 + \beta_2 \neq 0 \vee \beta_3 + \beta_4 \neq 0$. Note that in this paper, we will focus only on hypotheses containing linear combinations of regression coefficients. These combinations are specified by means of the $\boldsymbol{R}$-matrix and each part of the hypothesis can be expressed as a row in $\boldsymbol{R}$:

$$\boldsymbol{r}_1' = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \end{pmatrix}, \tag{7}$$

$$\boldsymbol{r}_2' = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \tag{8}$$

leading to the full constraint matrix:

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{9}$$

Then the hypothesis of interest can be expressed as $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}$ vs. $H_a : \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{0}$. Note that all kinds of hypotheses, including the single parameter case and comparisons of nested models, as discussed before, can be expressed by means of the $\boldsymbol{R}$-matrix.

In case our hypothesis of interest contains inequality constraints, like $H_a : \beta_1 + \beta_2 > 0 \vee \beta_3 + \beta_4 > 0$, $\boldsymbol{R}$ still looks the same, but we need to fit a model where we enforce the inequality constraints on the regression coefficients. This can be done by means of quadratic programming, for example using the subroutine `solve.QP()` of the R package quadprog (Turlach and Weingessel, 2019). It implements the dual method of Goldfarb and Idnani (1982, 1983). If we apply this method in the linear regression context, it has the following form (see "Data Sheet 1" in the Supplementary materials for further explanations):

$$min(-\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}) \qquad \text{with the constraints } \boldsymbol{R}\boldsymbol{\beta} \geq \boldsymbol{\beta}_0. \tag{10}$$

Note that all quantities based on an inequality constrained model are denoted by a tilde on top of them. Assume that the unconstrained estimates $\hat{\boldsymbol{\beta}}'$ are $(0.100 \quad -0.130 \quad 0.100 \quad -0.240 \quad 0.250)$, but the inequality constrained estimates $\tilde{\boldsymbol{\beta}}'$ may be $(0.110 \quad -0.110 \quad 0.120 \quad -0.230 \quad 0.240)$, where the estimates of β_0, β_3 and β_4 may also change slightly, even though they already satisfied the constraints in the unrestricted estimation. The restricted estimation will also lead to different residuals than the unrestricted estimation.

If our hypothesis of interest contains equality constraints, for example $H_a : \beta_1 + \beta_2 = 0 \vee \beta_3 + \beta_4 = 0$, the equality constrained estimates $\bar{\boldsymbol{\beta}}$ can also be found via quadratic programming. Note that here, H_a from informative hypothesis testing equals H_0 from classical null hypothesis testing. Similarly, all estimated quantities with a bar on top are both the quantities from the equality constrained fit in informative hypothesis testing and the quantities obtained based on H_0 in classical null hypothesis testing, which are in fact equality constrained estimates as well. The corresponding mean squared error terms for the inequality and equality constrained case are defined as follows:

$$\tilde{S}^2_{corrected} = \frac{\widetilde{RSS}}{n-k}, \tag{11}$$

$$\tilde{S}^2_{naive} = \frac{\widetilde{RSS}}{n}, \tag{12}$$

$$\bar{S}^2_{corrected} = \frac{\overline{RSS}}{n-(k-h)}, \tag{13}$$

$$\bar{S}^2_{naive} = \frac{\overline{RSS}}{n}, \tag{14}$$

where $\widetilde{RSS}$ is the residual sum of squares of the inequality constrained fit $\sum_{i=1}^{n} \tilde{e}_i^2$, where $\tilde{e}_i = y_i - \tilde{y}_i$ and $\tilde{y}_i$ are the model predicted values of the response variable. Furthermore, $\overline{RSS}$ is the residual sum of squares under the equality constrained fit $\sum_{i=1}^{n} \bar{e}_i^2$, where $\bar{e}_i = y_i - \bar{y}_i$ and $\bar{y}_i$ are the model predicted values of the response variable. Finally, h is the row rank of $\boldsymbol{R}$.

Similarly, we can define the unit information matrices of the inequality and equality constrained fits:

$$\tilde{\boldsymbol{I}}_1 = \frac{1}{n\,\tilde{S}^2_{corrected}}\boldsymbol{X}'\boldsymbol{X}, \tag{15}$$

$$\bar{\boldsymbol{I}}_1 = \frac{1}{n\,\bar{S}^2_{corrected}}\boldsymbol{X}'\boldsymbol{X}. \tag{16}$$

Note that $\boldsymbol{X}$ from the inequality constrained fit equals $\boldsymbol{X}$ from the unconstrained fit. The estimates $\hat{\boldsymbol{\beta}}, \tilde{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}}$ as well as the corresponding mean squared error terms and unit information matrices are used in the test statistics that are presented in the subsequent section.

Hypothesis testing

In order to give a broad overview about different test statistics, we present regular test statistics used in classical null hypothesis testing, as well as informative test statistics used in informative hypothesis testing. Note that an overview table containing all test statistics is provided at the end of each section. All test statistics can be applied in the setting of linear regression. "Data Sheet 2" in the Supplementary materials shows how these test statistics are implemented in R code.

Classical null hypothesis testing

The test statistics from classical null hypothesis testing that we will explain include the Wald test, the LRT, the Score test, the F-test as well as the t-test. The large sample test statistics, that is the Wald test, the LRT and the Score test, can be defined as follows Buse (1982):

$$Wald = n(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}), \tag{17}$$

$$LRT = -2\cdot[\ell(\bar{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})], \tag{18}$$

$$Score = \frac{1}{n}\boldsymbol{S}(\bar{\boldsymbol{\beta}})'\bar{\boldsymbol{I}}_1^{-1}\boldsymbol{S}(\bar{\boldsymbol{\beta}}), \tag{19}$$

where $\ell(\bar{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\bar{\boldsymbol{\beta}}$, $\ell(\hat{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{S}(\bar{\boldsymbol{\beta}}) = \frac{\partial}{\partial\bar{\boldsymbol{\beta}}}\ell(\bar{\boldsymbol{\beta}})$ is the score function evaluated at $\bar{\boldsymbol{\beta}}$. All three test statistics follow asymptotically a χ^2-distribution under the null hypothesis with $df = h$, if the model is correct.

Note that all three test statistics implicitly depend on S^2 in the information matrices (see Equation 6) and in the log-likelihoods. In the regression setting, since we always know what the residual degrees of freedom are, we can use $\hat{S}^2_{corrected}$ instead of $\hat{S}^2_{naive}$ to obtain the corrected instead of naive test statistic versions. That way, we can use the F-distribution with $df_1 = h, df_2 = n - p$ to obtain the p-values, which is more precise in small samples compared to the χ^2-distribution.

Note that the LRT, the Wald and the Score test are asymptotically equivalent. However, it has been shown that the values of the Wald test are always slightly larger than the values of the LRT, which in turn are always slightly larger than the values of the Score test (Buse, 1982, p. 157). Thus, using the same critical χ^2 value, the tests may have different power properties, which can be one aspect guiding the choice between them. Another aspect may be the time it takes to compute the three tests. For the Wald test, we need to fit the unconstrained model, whereas for the Score test, we need to fit the equality constrained model and for the LRT, we need to fit both the unconstrained and equality constrained model. In many cases, fitting the unconstrained model takes the least amount of time, which is why the Wald test is chosen often. However, in some cases, for example if the equality constrained model has a lot less parameters than the unconstrained model, it may be faster to fit the equality constrained model compared to the unconstrained model.

The F-test can be calculated as Seber and Lee (2012, p. 100):

$$F_{corrected} = \frac{\frac{1}{h}[\widehat{RSS} - \widehat{RSS}]}{\hat{S}^2_{corrected}}. \tag{20}$$

Another test statistic version results from using $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$, which we denote as F_{naive}. Seber and Lee (2012, p. 100) show that $F_{corrected}$ can be re-written to contain the unit information matrix:

$$F^{info}_{corrected} = \frac{n}{h}(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}), \tag{21}$$

where the superscript "info" refers to the information matrix. When $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$ is used in constructing the unit information matrix, we call this test statistic F^{info}_{naive}. If the model is specified correctly, $F_{corrected}$ follows an F-distribution with $df_1 = h, df_2 = n - k$ under the null hypothesis.

The one-sample t-test is defined as Allen(1997, p. 67):

$$t = \frac{\hat{\beta} - \bar{\beta}}{SE_{\hat{\beta}}}, \tag{22}$$

where $\bar{\beta}$ is the value of β under the null hypothesis and $SE_{\hat{\beta}}$ is the standard error of $\hat{\beta}$. Under the null hypothesis, t is t-distributed with $df = n - k$, if the model is correct. Note that if $h = 1$ the t- and F-statistic are related in a certain way, which is $t^2 = F$.

It is widely known that the one-sample t-test can be used for testing both two-sided hypotheses like $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ as well as one-sided hypotheses like $H_0 : \beta = 0$ against $H_a : \beta > 0$ or $H_a : \beta < 0$. The test statistic stays the same in both cases, but the p-value is computed differently. That is, when testing a two-sided hypothesis, half of the significance level is allocated to each side of the t-distribution, whereas when testing a one-sided hypothesis, all of it is allocated to one side of the t-distribution. That means that the cut-off levels, denoting from

TABLE 1 Overview of all presented regular test statistics.

Regular test statistics	Formula
$LRT_{naive/corrected}$	$-2 \cdot [\ell(\bar{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})]$
$Wald_{naive/corrected}$	$n(\boldsymbol{R}\hat{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}})$
$Score_{naive/corrected}$	$\frac{1}{n}\boldsymbol{S}(\bar{\boldsymbol{\beta}})'\bar{\boldsymbol{I}}_1^{-1}\boldsymbol{S}(\bar{\boldsymbol{\beta}})$
$F_{naive/corrected}$	$\frac{\frac{1}{p}[\bar{RSS}-\hat{RSS}]}{\hat{S}^2_{naive/corrected}}$
t	$\frac{\hat{\beta}-\bar{\beta}}{SE_{\bar{\beta}}}$

which point on the t-statistic can be considered to be significant, change. The two-sided p-value, which is the default output of most statistical software, simply adds up the probabilities of the negative and positive version of the observed t-value (t_{obs}), independently of whether it was in fact positive or negative:

$$\begin{aligned} p_{two-sided} &= 2 \cdot P(t > |t_{obs}|) \\ &= P(t > t_{obs}) + P(t < -t_{obs}). \end{aligned} \tag{23}$$

Since the t-distribution is symmetric, $P(t > t_{obs})$ is the same as $P(t < -t_{obs})$. When we are interested in the one-sided p-value and $H_a : \beta > 0$, the p-value is obtained as:

$$p_{one-sided} = P(t > t_{obs}), \tag{24}$$

whereas if $H_a : \beta < 0$, the p-value is obtained as:

$$p_{one-sided} = P(t < t_{obs}). \tag{25}$$

Note that in case the obtained t-value is a positive number and we are interested in $H_a : \beta > 0$ or in case t is a negative number and we are interested in $H_a : \beta < 0$, the one-sided p-value can be obtained by dividing the two-sided p-value by 2.

In summary, the t-statistic is a special case, since this statistic from the classical null hypothesis testing framework can be used for testing an informative hypothesis, as long as the hypothesis only contains one parameter. If we are interested in more than one parameter, we can no longer use the t-statistic, but have to use an informative test statistic. Table 1 shows an overview about all presented regular test statistics.

Informative hypothesis testing

Informative test statistics are often a modified version of the regular test statistics. In case the model is correct, the large sample informative test statistics, including the LRT, the Wald test, the Score test and the D-statistic, asymptotically follow a $\bar{\chi}^2$-distribution under the null hypothesis, which is a mixture of χ^2-distributions. The small sample informative test statistic, that is the $\bar{F}$-statistic, follows an $\bar{F}$-distribution under the null hypothesis, if the model is correctly specified. The $\bar{F}$-distribution is a mixture of F-distributions. Note that similar to classical null hypothesis testing, we can use the corrected instead of naive mean squared error to obtain the large sample test statistics. In that way, we can calculate the p-values by means of the $\bar{F}$-distribution instead of the $\bar{\chi}^2$-distribution to obtain more precise results in small sample sizes.

The $LRT_{corrected}$ test statistic can be calculated as follows Silvapulle and Sen (2005, p. 157):

$$LRT_{corrected} = -2 \cdot [\ell(\bar{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})], \tag{26}$$

where $\ell(\bar{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\bar{\boldsymbol{\beta}}$ and $\ell(\tilde{\boldsymbol{\beta}})$ is the log-likelihood evaluated at $\tilde{\boldsymbol{\beta}}$. $\ell(\bar{\boldsymbol{\beta}})$ has been calculated using $\bar{S}^2_{corrected}$ and $\ell(\tilde{\boldsymbol{\beta}})$ has been calculated using $\tilde{S}^2_{corrected}$. If $\bar{S}^2_{naive}$ and $\tilde{S}^2_{naive}$ were used instead, we would obtain LRT_{naive}.

The Wald statistic can be found in Silvapulle and Sen (2005, p. 154):

$$Wald^{info}_{corrected} = \frac{n}{\tilde{S}^2_{corrected}}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}), \tag{27}$$

where $\boldsymbol{W} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X}$. The Wald version where we use $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$ is called $Wald^{info}_{naive}$. Both versions implicitly contain $\hat{\boldsymbol{I}}_1$ (see Equation 6), which can also be replaced by $\tilde{\boldsymbol{I}}_1$. Note that $Wald^{info}_{naive}$ will give different results, especially in small sample sizes, due to the missing correction. Assuming $VCOV(\hat{\boldsymbol{\beta}})$ is defined as in Equation 5, we can re-write the Wald statistic as:

$$Wald^{VCOV} = [\boldsymbol{R}\tilde{\boldsymbol{\beta}}]'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}[\boldsymbol{R}\tilde{\boldsymbol{\beta}}], \tag{28}$$

which is identical to $Wald^{info}_{corrected}$. Note that we can also replace $VCOV(\hat{\boldsymbol{\beta}})$ by a more robust sandwich-estimator, which is not commonly done in the applied literature.

The D-statistic is calculated as follows (Silvapulle and Sen, 2005, p. 159):

$$D_{corrected} = \frac{2 \cdot n}{\tilde{S}^2_{corrected}}[d(\bar{\boldsymbol{\beta}}) - d(\tilde{\boldsymbol{\beta}})], \tag{29}$$

where $d(\bar{\boldsymbol{\beta}})$ and $d(\tilde{\boldsymbol{\beta}})$ are the values of the following two functions at their solutions (see "Data Sheet 3" in the Supplementary materials for further information):

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad \text{under the constraint } \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{0}, \tag{30}$$

$$f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad \text{under the constraint } \boldsymbol{R}\boldsymbol{\beta} \geq \boldsymbol{0}. \tag{31}$$

When minimizing these functions, we treat $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{W}$ as known constants. Note that in the regression case, $D_{corrected}$ is identical to $Wald^{info}_{corrected}$ and $Wald^{VCOV}$, as long as $\hat{S}^2_{corrected}$ is used. In contrast, if we switch to using $\hat{S}^2_{naive}$, we obtain D_{naive}, in which case $D_{naive} = Wald^{info}_{naive}$.

The $\bar{F}$-statistic can be found in (Silvapulle and Sen, 2005, p. 29):

$$\bar{F}_{corrected} = \frac{\overline{RSS} - \widetilde{RSS}}{\hat{S}^2_{corrected}}. \quad (32)$$

According to Silvapulle and Sen (2005, p. 29), including the constant $\frac{1}{h}$ from the regular F-statistic in the $\bar{F}$-statistic is not necessary, as it does not affect the results. Again, when using $\hat{S}^2_{naive}$ instead of $\hat{S}^2_{corrected}$, we obtain $\bar{F}_{naive}$. We can re-write the $\bar{F}$-statistic similarly to how we re-wrote the F-statistic. Assuming that we use $\hat{S}^2_{corrected}$ to compute the unit information matrix, we obtain:

$$\bar{F}^{info}_{corrected} = n(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}}). \quad (33)$$

Again, $\hat{\boldsymbol{I}}_1$ can be replaced by $\tilde{\boldsymbol{I}}_1$.

There are various versions of the Score statistic. $Score^{U}_{corrected}$ can be found in Silvapulle and Sen (2005, p. 159):

$$Score^{U}_{corrected} = \frac{1}{n \cdot \hat{S}^2_{corrected}} \boldsymbol{U}'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{U}, \quad (34)$$

where $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{W}^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$. When using $\hat{S}^2_{naive}$ as compared to $\hat{S}^2_{corrected}$, we obtain $Score^{U}_{naive}$. Another version of the Score statistic, $Score^{null-info}_{corrected}$, is defined as follows Silvapulle and Silvapulle (1995, p. 342):

$$Score^{null-info}_{corrected} = \frac{1}{n}[\boldsymbol{S}(\bar{\boldsymbol{\beta}}) - \boldsymbol{S}(\tilde{\boldsymbol{\beta}})]'\bar{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\bar{\boldsymbol{\beta}}) - \boldsymbol{S}(\tilde{\boldsymbol{\beta}})], \quad (35)$$

where $\bar{\boldsymbol{I}}_1$ has been calculated by means of $\bar{S}^2_{corrected}$ (see Equation 13). In contrast, if we use $\bar{S}^2_{naive}$, we obtain $Score^{null-info}_{naive}$.

Furthermore, $Score^{info}_{corrected}$ can be calculated as Silvapulle and Sen (2005, p. 166):

$$Score^{info}_{corrected} = \frac{1}{n}\boldsymbol{P}'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}\boldsymbol{P}, \quad (36)$$

where $\boldsymbol{P} = \boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$ and $\hat{\boldsymbol{I}}_1$ is calculated using $\hat{S}^2_{corrected}$ and can be replaced by either $\tilde{\boldsymbol{I}}_1$ or $\bar{\boldsymbol{I}}_1$. If we use $\hat{S}^2_{naive}$ to calculate $\hat{\boldsymbol{I}}_1$, we obtain $Score^{info}_{naive}$. Silvapulle and Sen (2005, p. 166) mention another way to express $Score^{info}_{corrected}$:

$$Score^{info,Robertson}_{corrected} = \frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})], \quad (37)$$

where the superscript "Robertson" indicates that this is the version defined by Robertson et al. (1988), $\hat{\boldsymbol{I}}_1$ is calculated using $\hat{S}^2_{corrected}$ and can be replaced by either $\tilde{\boldsymbol{I}}_1$ or $\bar{\boldsymbol{I}}_1$. Assuming that $VCOV(\hat{\boldsymbol{\beta}})$ is defined as in Equation 5, $Score^{info}_{corrected}$ can be re-written as:

$$Score^{VCOV} = \boldsymbol{V}'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}\boldsymbol{V}, \quad (38)$$

TABLE 2 Overview of all presented informative test statistics.

Informative test statistics	Formulas
$LRT_{naive/corrected}$	$-2 \cdot [\ell(\bar{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})]$
$Wald^{info}_{naive}$	$\frac{n}{\hat{S}^2_{naive}}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})$
$Wald^{info}_{corrected} = Wald^{VCOV}$	$\frac{n}{\hat{S}^2_{corrected}}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})$ $= [\boldsymbol{R}\tilde{\boldsymbol{\beta}}]'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}[\boldsymbol{R}\tilde{\boldsymbol{\beta}}]$
$D_{naive/corrected}$	$\frac{2 \cdot n}{\hat{S}^2_{naive/corrected}}[d(\bar{\boldsymbol{\beta}}) - d(\tilde{\boldsymbol{\beta}})]$
$\bar{F}_{naive}$	$\frac{\overline{RSS} - \widetilde{RSS}}{\hat{S}^2_{naive}}$
$\bar{F}_{corrected} = \bar{F}^{info}_{corrected}$	$\frac{\overline{RSS} - \widetilde{RSS}}{\hat{S}^2_{corrected}}$ $= n(\boldsymbol{R}\tilde{\boldsymbol{\beta}})'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}(\boldsymbol{R}\tilde{\boldsymbol{\beta}})$
$Score^{U}_{naive/corrected}$	$\frac{1}{n \cdot \hat{S}^2_{naive/corrected}}\boldsymbol{U}'(\boldsymbol{R}\boldsymbol{W}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{U}$
$Score^{null-info}_{naive/corrected}$	$\frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\bar{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$
$Score^{info}_{naive} = Score^{info,Robertson}_{naive}$	$\frac{1}{n}\boldsymbol{P}'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}\boldsymbol{P}$ $= \frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$
$Score^{info}_{corrected} = Score^{info,Robertson}_{corrected}$ $= Score^{VCOV}$	$\frac{1}{n}\boldsymbol{P}'(\boldsymbol{R}\hat{\boldsymbol{I}}_1^{-1}\boldsymbol{R}')^{-1}\boldsymbol{P}$ $= \frac{1}{n}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]'\hat{\boldsymbol{I}}_1^{-1}[\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$ $= \boldsymbol{V}'[\boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ \boldsymbol{R}']^{-1}\boldsymbol{V}$

where $\boldsymbol{V} = \boldsymbol{R}\ VCOV(\hat{\boldsymbol{\beta}})\ [\boldsymbol{S}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{S}(\bar{\boldsymbol{\beta}})]$, again allowing for a more robust sandwich-estimator of $VCOV(\hat{\boldsymbol{\beta}})$ to be inserted. Table 2 gives an overview about all the informative test statistics that were presented.

P-values

There are two approaches for calculating the *p*-value of informative test statistics (Silvapulle and Sen, 2005). In this paper, we use the approach where we first calculate the weights of the respective mixture distribution $(\bar{\chi}^2, \bar{F})$. Note that the sum of the weights from 0 to q is one, where q is the rank of $\boldsymbol{X}$ under the null hypothesis.

If the residuals of our data are normally distributed, we can use the multivariate normal probability function as well as the `ic.weight()` function of the R package ic.infer (Grömping, 2010) to compute the weights. These calculations are also implemented in the R package restriktor (Vanbrabant, 2020). Once we have computed the weights, the *p*-values of the observed $\bar{\chi}^2$-value ($\bar{\chi}^2_{obs}$) and of the observed $\bar{F}$-value ($\bar{F}_{obs}$) are obtained as follows Silvapulle and Sen (2005, pp. 86 and 99):

$$\Pr(\bar{\chi}^2 \geq \bar{\chi}^2_{obs}) = \sum_{i=0}^{q} w_i(H_0, H_a)\Pr[(h - q + i)\chi^2_{h-q+i} \geq \bar{\chi}^2_{obs}], \quad (39)$$

$$\Pr(\bar{F} \geq \bar{F}_{obs}) = \sum_{i=0}^{q} w_i(H_0, H_a)\Pr[(h - q + i)F_{h-q+i,n-p} \geq \bar{F}_{obs}]. \quad (40)$$

TABLE 3 Type I error rates when using R_1 and applying the test statistics as outlined in the referenced books.

n	$LRT_{corr.}$	$LRT_{restr.}$	$Wald^{info}_{naive}$	$Wald^{info}_{corr.}$ $Wald^{VCOV}$ $D_{corr.}$	$Score^{U}_{corr.}$	$Score^{null-info}_{corr.}$	$Score^{null-info}_{restr.}$	$Score^{info}_{corr.}$ $Score^{VCOV}$	$\bar{F}_{corr.}$ $\bar{F}_{restr.}$	$F_{corr.}$	$t_{one-s.}$	$t_{two-s.}$
10000	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.047	0.050	0.047	0.050
2000	0.054	0.054	0.054	0.054	0.057	0.054	0.054	0.054	0.054	0.060	0.054	0.060
1000	0.057	0.057	0.058	0.057	0.058	0.057	0.057	0.057	0.057	0.067	0.057	0.067
500	0.058	0.058	0.058	0.055	0.053	0.055	0.055	0.055	0.055	0.049	0.055	0.049
100	0.054	0.054	0.056	0.051	0.054	0.050	0.048	0.048	0.049	0.043	0.049	0.043
50	0.057	0.058	0.060	0.054	**0.066**	0.051	0.044	0.044	0.049	0.044	0.049	0.044
25	**0.074**	**0.074**	**0.092**	**0.066**	**0.089**	0.057	0.047	0.045	0.057	0.057	0.057	0.057
10	**0.125**	**0.112**	**0.186**	**0.098**	**0.169**	0.054	0.002	0.000	0.054	**0.061**	0.054	**0.061**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $LRT_{restriktor}$ as $LRT_{restr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$, $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$, $Score^{null-info}_{corrected}$ as $Score^{null-info}_{corr.}$, $Score^{null-info}_{restriktor}$ as $Score^{null-info}_{restr.}$, $Score^{info}_{corrected}$ as $Score^{info}_{corr.}$, $\bar{F}_{corrected}$ as $\bar{F}_{corr.}$, $F_{corrected}$ as $F_{corr.}$, $t_{one-sided}$ as $t_{one-s.}$ and $t_{two-sided}$ as $t_{two-s.}$. Bold values are above 0.06 and underlined values are below 0.04.

TABLE 4 Type I error rates when using R_2 and applying the test statistics as outlined in the referenced books.

n	$LRT_{corr.}$	$LRT_{restr.}$	$Wald^{info}_{naive}$	$Wald^{info}_{corr.}$ $Wald^{VCOV}$ $D_{corr.}$	$Score^{U}_{corr.}$	$Score^{null-info}_{corr.}$	$Score^{null-info}_{restr.}$	$Score^{info}_{corr.}$ $Score^{VCOV}$	$\bar{F}_{corr.}$ $\bar{F}_{restr.}$	$F_{corr.}$
10000	0.052	0.052	0.052	0.052	0.049	0.052	0.052	0.052	0.052	0.049
2000	0.048	0.048	0.050	0.048	0.052	0.048	0.048	0.048	0.048	0.046
1000	0.051	0.051	0.051	0.051	0.052	0.051	0.047	0.047	0.051	0.058
500	0.059	0.059	**0.062**	0.060	0.059	0.059	0.057	0.057	0.059	0.048
100	0.057	0.056	**0.070**	**0.061**	**0.078**	0.055	0.053	0.051	0.056	0.058
50	0.051	0.046	**0.090**	0.060	**0.099**	0.044	0.039	0.035	0.048	0.055
25	**0.068**	0.055	**0.135**	**0.083**	**0.119**	0.052	0.027	0.010	**0.064**	0.055
10	**0.069**	0.011	**0.416**	**0.163**	**0.334**	0.024	0.001	0.000	0.054	**0.061**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $LRT_{restriktor}$ as $LRT_{restr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$, $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$, $Score^{null-info}_{corrected}$ as $Score^{null-info}_{corr.}$, $Score^{null-info}_{restriktor}$ as $Score^{null-info}_{restr.}$, $Score^{info}_{corrected}$ as $Score^{info}_{corr.}$, $\bar{F}_{corrected}$ as $\bar{F}_{corr.}$, $F_{corrected}$ as $F_{corr.}$. Bold values are above 0.06 and underlined values are below 0.04.

It can be expected that the p-values are very similar, irrespective of whether they are calculated based on the $\bar{\chi}^2$- or $\bar{F}$-distribution, as long as sample sizes are large. However, for small sample sizes, the $\bar{F}$-distribution should yield more accurate results.

Simulation studies

We conducted several simulation studies to examine the impact of different conditions on the performance of the presented test statistics in terms of type I and type II error rates. We were interested in the effects of sample and effect sizes, the number of regression parameters considered in H_a as well as the distribution used for calculating the p-values. Our main motivation was to provide a reference framework for applied researchers who wish to test informative hypotheses, helping them to chose the optimal test statistic(s) in the present situation.

Design

We generated a design matrix $\boldsymbol{X}$, including data for five regression coefficients $\boldsymbol{\beta}' = (\beta_1\ \beta_2\ \beta_3\ \beta_4\ \beta_5)$ and considered effect sizes of $f^2 = 0.02$ (small), $f^2 = 0.10$ (medium) and $f^2 = 0.35$ (large) and sample sizes of $10, 25, 50, 100, 500, 1000, 2000,$ and 10000. For examining the type I error rate, we generated a random outcome Y, whereas for examining the type II error rate, we fixed all βs to 0.1 and generated y with a random error term that was specific for the effect size used. Since $f^2 = \frac{R^2}{1-R^2}$, where R^2 is the determination coefficient, we can calculate the error terms of y by plugging in the f^2-specific value of R^2 in

$$S^2_y = [\boldsymbol{\beta}\ Cor(\boldsymbol{X})\ \boldsymbol{\beta}] \times \frac{1-R^2}{R^2}, \tag{41}$$

where $Cor(\boldsymbol{X})$ is the correlation matrix of the design matrix $\boldsymbol{X}$. The number of replications was 1000.

TABLE 5 Type I error rates when using R_1, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}, \bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

n	$LRT_{corr.}$	$Wald^{info}_{corr.}$ $D_{corr.}$	$Score^{U}_{corr.}$
10000	0.047	0.047	0.047
2000	0.054	0.054	0.057
1000	0.057	0.057	0.058
500	0.055	0.055	0.053
100	0.052	0.049	0.052
50	0.055	0.049	**0.065**
25	**0.067**	0.057	**0.080**
10	**0.084**	0.054	**0.126**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$ and $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$. Bold values are above 0.06 and underlined values are below 0.04.

TABLE 6 Type I error rates when using R_2, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}, \bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

n	$LRT_{corr.}$	$Wald^{info}_{corr.}$ $D_{corr.}$	$Score^{U}_{corr.}$
10000	0.052	0.052	0.049
2000	0.048	0.048	0.051
1000	0.050	0.051	0.051
500	0.059	0.059	0.058
100	0.055	0.056	**0.072**
50	0.043	0.048	**0.085**
25	0.042	**0.064**	**0.097**
10	0.006	0.054	**0.161**

The test statistics are abbreviated as follows: $LRT_{corrected}$ as $LRT_{corr.}$, $Wald^{info}_{corrected}$ as $Wald^{info}_{corr.}$, $D_{corrected}$ as $D_{corr.}$ and $Score^{U}_{corrected}$ as $Score^{U}_{corr.}$. Bold values are above 0.06 and underlined values are below 0.04.

We considered two different kinds of $\boldsymbol{R}$ matrices, where the first one was defined as follows:

$$\boldsymbol{R}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{42}$$

This represents the hypothesis that only β_1 is greater than zero: $H_a : \beta_1 > 0$. The second $\boldsymbol{R}$ matrix was defined as:

$$\boldsymbol{R}_2 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \tag{43}$$

stating that at least one of the regression coefficients, except the intercept, are greater than zero: $H_a : \beta_1 > 0 \vee \beta_2 > 0 \vee \beta_3 > 0 \vee \beta_4 > 0 \vee \beta_5 > 0$.

To compute the test statistics, we used $\hat{S}^2_{naive}$ and $\hat{S}^2_{corrected}$ as well as $\tilde{S}^2_{naive}, \tilde{S}^2_{corrected}, \bar{S}^2_{naive}$ and $\bar{S}^2_{corrected}$ and to compute the p-values, we used the $\bar{\chi}^2$- as well as the $\bar{F}$-distribution. In addition to the manual calculation of the test statistics, we also included the test statistics as reported by the R package restriktor.

Type I results

Test statistics were first applied the way they are presented in the referenced literature. That is, $Wald^{info}_{naive}$ makes use of $\hat{S}^2_{naive}$, whereas all other test statistics make use of $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}, \bar{S}^2_{corrected}$). For calculating the p-values, the $\bar{\chi}^2$-distribution is used for $LRT_{corrected}$, $Wald^{info}_{naive}$, $Wald^{info}_{corrected}$, $Wald^{VCOV}$, $D_{corrected}$, $Score^{U}_{corrected}$, $Score^{null-info}_{corrected}$, $Score^{info}_{corrected}$ and $Score^{VCOV}$. The $\bar{F}$-distribution is used for calculating the p-values for the $\bar{F}$-statistic, the F-distribution is used for calculating the p-values for the F-statistic and the t-distribution is used for calculating the p-values for the t-statistic. Note that restriktor always uses $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}, \bar{S}^2_{corrected}$) for all available test statistics and always calculates the p-value based on the $\bar{F}$-distribution. Tables 3, 4 show the results.

We can observe that when using $\boldsymbol{R}_1$ (see Table 3), that is when testing a hypothesis concerning only one regression parameter, type I error rates are identical between F and $t_{two-sided}$ as well as between $\bar{F}$ and $t_{one-sided}$, showing the link between classical null hypothesis testing and informative hypothesis testing. When using $\boldsymbol{R}_2$ (see Table 4), that is when testing a hypothesis concerning multiple regression parameters, problems with type I error rates seem to occur earlier as compared to when using $\boldsymbol{R}_1$. More specifically, problematic type I error rates occur as early as with $n = 500$ or $n = 100$ when using $\boldsymbol{R}_2$, but only start occurring with $n = 50$ or $n = 25$ when using $\boldsymbol{R}_1$. Apart from that, $Score^{U}_{corrected}$ and $Wald^{info}_{naive}$ show the highest type I error rates for both $\boldsymbol{R}$ matrices, whereas $\bar{F}$ and $\bar{F}_{restriktor}$ show the most appropriate type I error rates for both $\boldsymbol{R}$ matrices. This is because the $\bar{F}$-distribution is more precise in small sample sizes as compared to the $\bar{\chi}^2$-distribution.

When using the $\bar{F}$-distribution instead of the $\bar{\chi}^2$-distribution when calculating the p-value for all test statistics, type I error rates are closer to the nominal level when sample sizes get smaller. This can be seen in Tables 5, 6 where a selection of test statistics are shown.

Furthermore, it can be observed that when using $\boldsymbol{R}_1$, type I error rates increase when using $LRT_{corrected}$ and $Score^{U}_{corrected}$ and $n = 10$ in contrast to $n = 25$. The same can only be observed for $Score^{U}_{corrected}$ when using $\boldsymbol{R}_2$, but not for $LRT_{corrected}$, where the type I error rate decreases quite substantially instead.

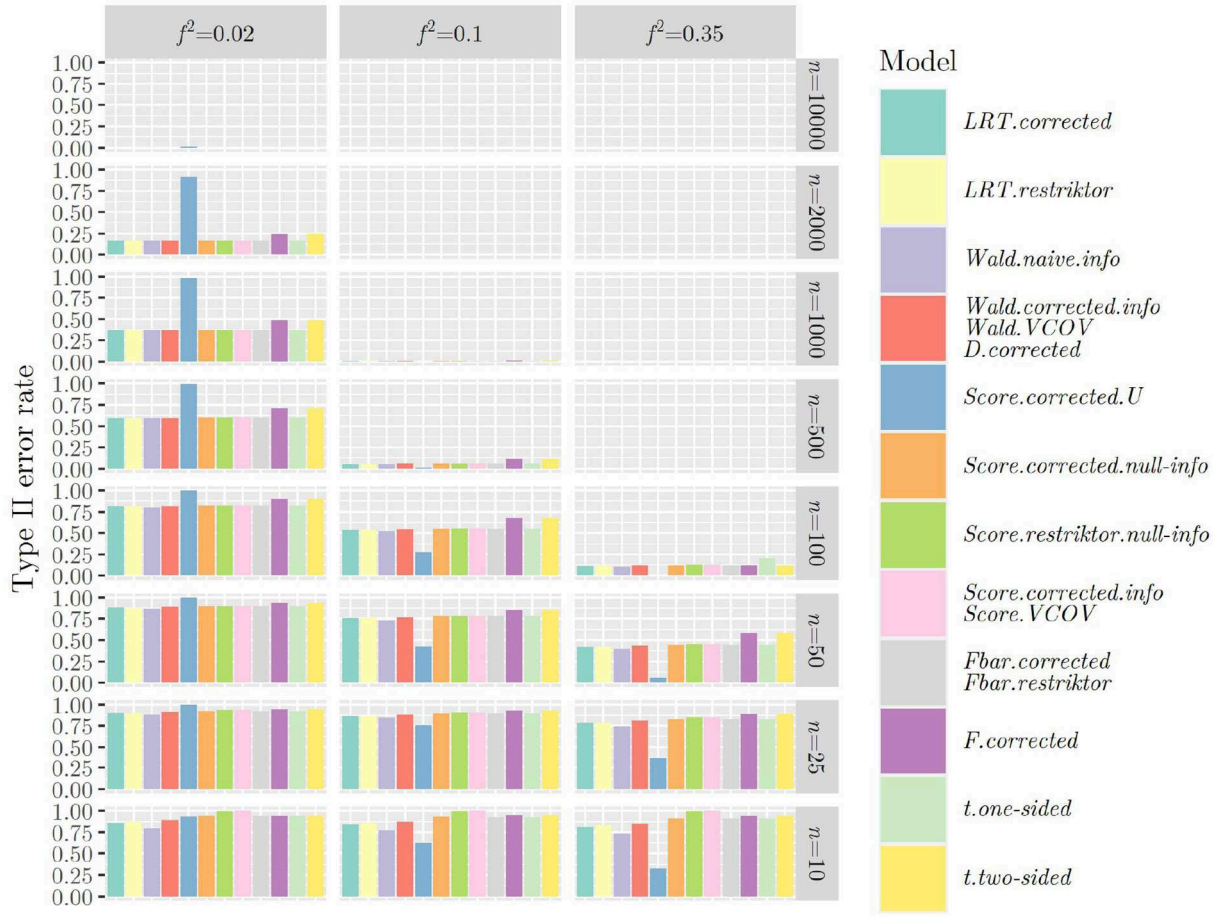

FIGURE 1
Type II error rates when using $\boldsymbol{R}_1$ and applying the test statistics as outlined in the referenced books.

More results can be found in "Data Sheet 4" in the Supplementary materials.

Type II results

Figures 1, 2 show the type II error rates when applying the test statistics as in the referenced books.

Once more, we can observe that when using $\boldsymbol{R}_1$ (see Figure 1), that is when testing a hypothesis concerning only one regression parameter, type II error rates are identical between F and $t_{two-sided}$ as well as between $\bar{F}$ and $t_{one-sided}$, showing the link between classical null hypothesis testing and informative hypothesis testing. When using $\boldsymbol{R}_2$ (see Figure 2), that is when testing a hypothesis concerning multiple regression parameters, problems with type II error rates seem to occur later (in terms of sample size) as compared to when using $\boldsymbol{R}_1$. This was the other way around regarding the type I error rate and it demonstrates the nature of the relationship between type I and type II error rates: If one goes down, the other one goes up and vice versa.

The same mechanism can be observed when using the $\bar{F}$-distribution instead of the $\bar{\chi}^2$-distribution when calculating the p-value for all test statistics (Figures 3, 4): Type II error rates are increased in small sample sizes, since type I error rates had improved, that is, decreased. Again, further results can be found in "Data Sheet 5" in the Supplementary materials.

Discussion

In this paper, we gave an overview of a large number of different informative test statistics, including their different versions. Furthermore, we clarified how those test statistics perform in terms of type I and type II error rates under different conditions by means of simulation studies in the context of linear regression. We considered varying sample and effect sizes as well as two different constraint matrices, where one

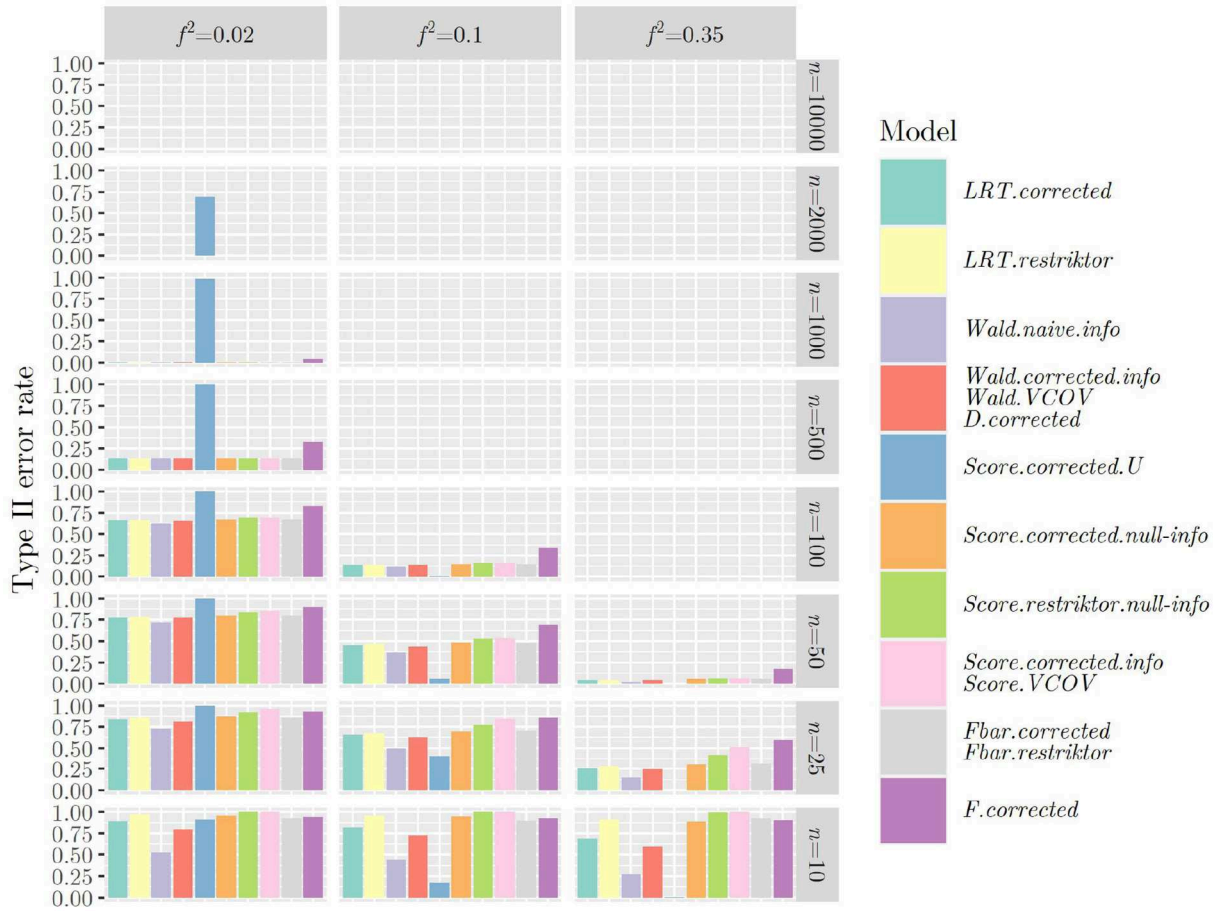

FIGURE 2
Type II error rates when using R_2 and applying the test statistics as outlined in the referenced books.

specified a hypothesis about one parameter and the other one specified a hypothesis about multiple parameters. Moreover, we considered the naive and corrected mean squared errors of the unconstrained, inequality and equality constrained models as part of the test statistics as well as the $\bar{\chi}^2$- and $\bar{F}$-distribution to calculate the *p*-values.

Based on our findings, the following recommendations can be made. Considering the time it takes to compute the informative test statistics, both the Wald and the $\bar{F}$-test versions are favorable, since they only need fitting of the inequality constrained model to obtain $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{I}}_1$. Even if we do not use $\tilde{\boldsymbol{I}}_1$ but use $\hat{\boldsymbol{I}}_1$ instead, the increase in time is small in the context of linear regression. The Score test and the LRT versions are less favorable, since they require fitting both the inequality constrained as well as the equality constrained model to obtain $\tilde{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\beta}}$, as well as the respective unit information matrices or log-likelihoods.

The *D*-statistic versions only require fitting the unconstrained model to obtain $\hat{\boldsymbol{\beta}}$. However, we then additionally need to compute the two functions $d(\bar{\boldsymbol{\beta}})$ and $d(\tilde{\boldsymbol{\beta}})$, which is as time-consuming as fitting the inequality constrained model. Thus, there is no advantage of using the *D*-statistic versions over the Wald and the $\bar{F}$-test versions in the context of linear regression. However, if the regression model was non-linear, computing the two functions would be significantly less computationally expensive than fitting the inequality constrained model.

Moreover, we recommend using the corrected mean squared error versions in the test statistics as well as using the $\bar{F}$-distribution for calculating the *p*-values, if sample sizes are small. This seems to keep type I error rates closer to the nominal level compared to using the naive mean squared error versions and using the $\bar{\chi}^2$-distribution for calculating the *p*-value. An additional interesting finding was that the relationship between LRT, Wald and Score test values that has been found in the unconstrained context also holds in the constrained context. That is, Wald test values are always slightly larger than LRT

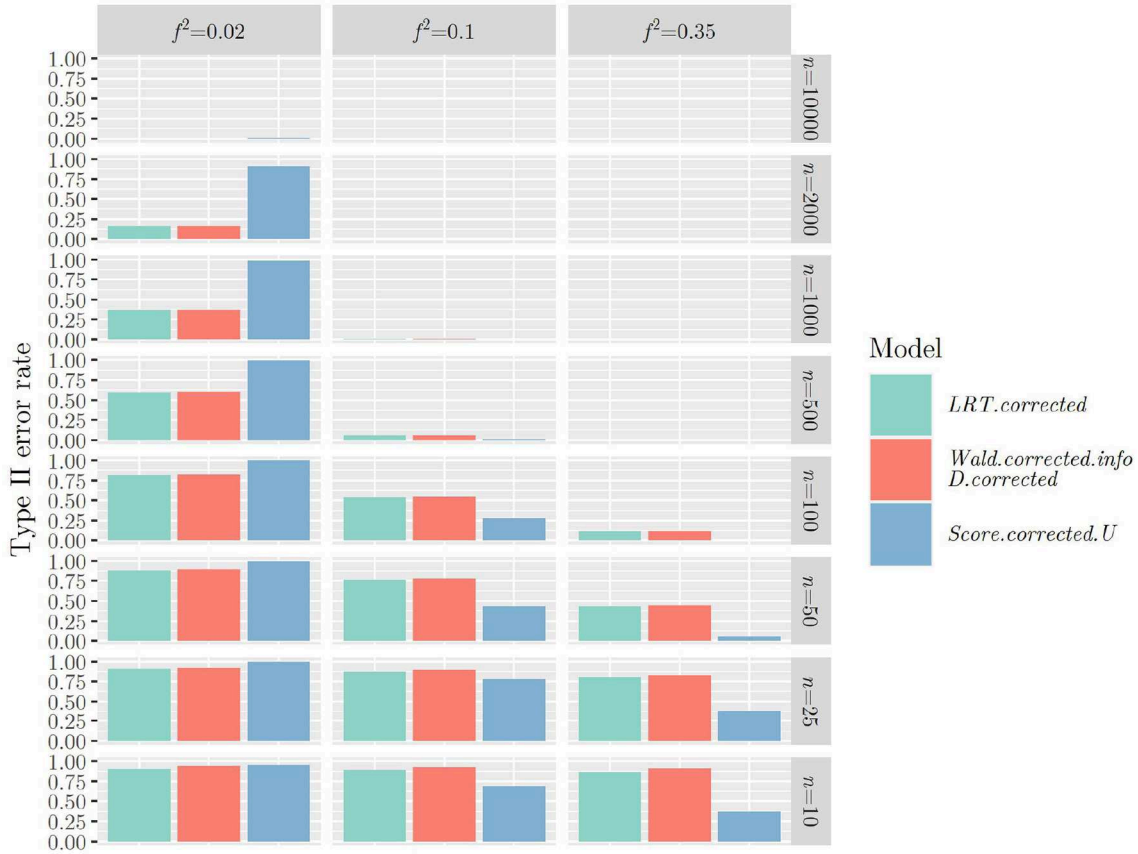

FIGURE 3
Type II error rates when using $\boldsymbol{R}_1$, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

values, which in turn are always slightly larger than Score test values.

The limitations of our simulation studies include the following aspects. We treated all variables as manifest, even though variables of interest in the social and behavioral sciences are often latent in nature. Furthermore, we solely generated normal data despite the fact that violations against the normality assumption occur regularly. Moreover, we used orthogonal predictors without interactions albeit this is rarely the case in the social and behavioral sciences. And lastly, we only included the regular versions of the standard errors and the variance-covariance matrix. Future research should thus repeat the simulation studies in the context of Structural Equation Modeling (SEM) to take into account latent variables. Furthermore, the impact of non-normal data as well as correlated predictors with interactions and using the robust versions of the standard errors and the variance-covariance matrix should be examined. It may be that under these conditions, type I and type II error rates deviate from the results presented in this paper. Moreover, the properties of informative test statistics, especially concerning the *D*-statistic, should also be investigated in the context of non-linear models.

Finally, research in the social and behavioral sciences is often not only interested in inference concerning regression coefficients, but also regarding effects of interest. These effects may be average or conditional treatment effects, which are defined as a linear or non-linear combination of regression coefficients. The EffectLiteR approach (Mayer et al., 2016) provides a framework and R package for the estimation of average and conditional effects of a discrete treatment variable on a continuous outcome variable, conditioning on categorical and continuous covariates. Keck et al. (2021) already demonstrated how to integrate informative hypothesis testing into the EffectLiteR framework in the context of linear regression. The present paper provides interested readers who want to apply informative hypothesis testing concerning regression coefficients or effects of interest with practical information regarding test statistics as well as type I and type II error rates.

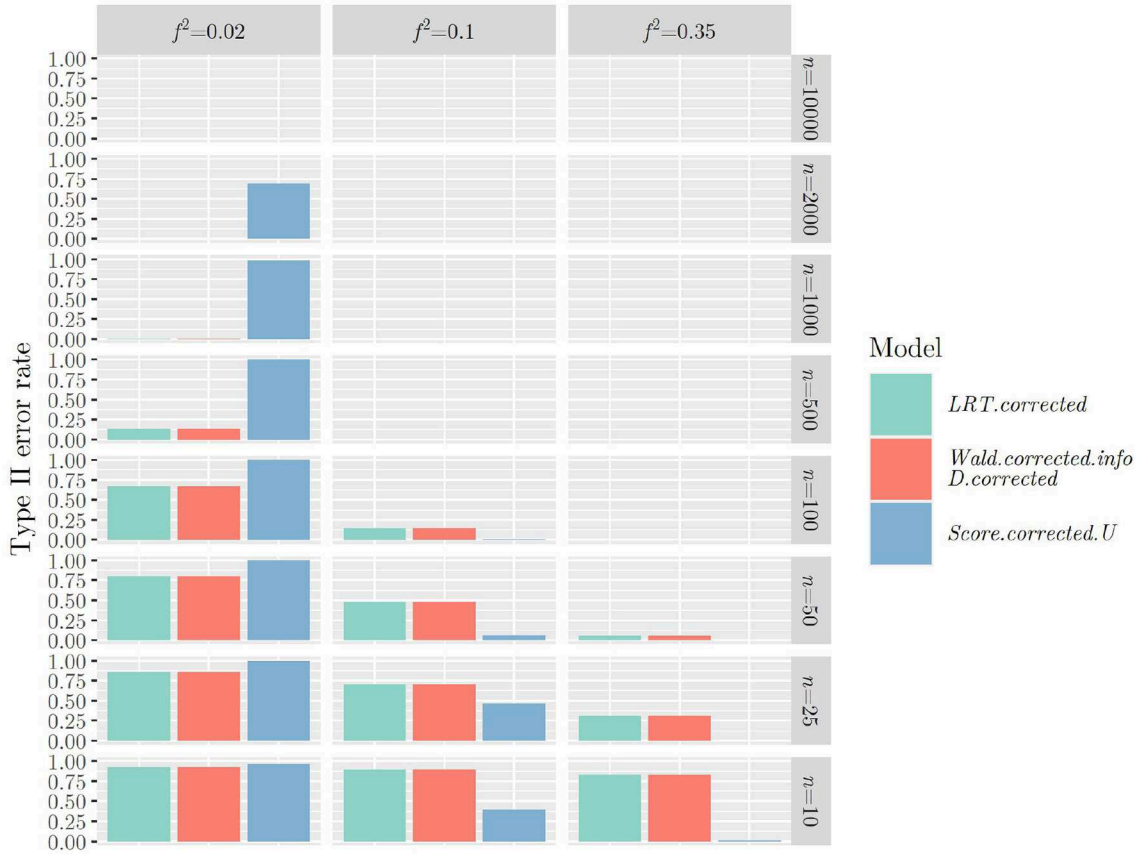

FIGURE 4
Type II error rates when using $\boldsymbol{R}_2$, $\hat{S}^2_{corrected}$ (or $\tilde{S}^2_{corrected}$, $\bar{S}^2_{corrected}$) and the $\bar{F}$-distribution for calculating the p-value.

Data availability statement

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

Author contributions

CK and YR contributed to conception and design of the study. CK performed the statistical analysis and wrote the first draft of the manuscript. CK, YR, and AM wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work has been supported by the Research Foundation Flanders (FWO, Grant No. G002819N to YR and AM).

Conflict of interest

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.899165/full#supplementary-material

References

Allen, M. P. (1997). *Understanding Regression Analysis*. Boston, MA: Springer.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions*. New York, NY: Wiley.

Buse, A. (1982). The likelihood ratio, Wald and Lagrange multiplier tests: An expository note. *Am. Stat.* 36, 153–157. doi: 10.1080/00031305.1982.10482817

Goldfarb, D., and Idnani, A. (1982). *Dual and Primal-Dual Methods for Solving Strictly Convex Quadratic Programs*, Berlin: Springer.

Goldfarb, D., and Idnani, A. (1983). A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program.* 27, 1–33. doi: 10.1007/BF02591962

Grömping, U. (2010). Inference with linear equality and inequality constraints using R: The package ic.infer. *J. Stat. Softw.* 33, 1–31. doi: 10.18637/jss.v033.i10

Gu, X., Hoijtink, H., Mulder, J., Van Lissa, C. J., Van Zundert, C., Jones, J., et al. (2020). *Bain: Bayes factors for informative hypotheses*. R package version 0.2.4.

Hoijtink, H. (2012). *Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists*. Boca Raton, FL: Chapman & Hall/CRC.

Hoijtink, H., Klugkist, I., and Boelen, P. A. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York, NY: Springer.

Huber, P. (1967). "The behavior of maximum likelihood estimates under nonstandard conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, Statistics* (Berkeley, CA: University of California Press), 221–233.

Keck, C., Mayer, A., and Rosseel, Y. (2021). Integrating informative hypotheses into the EffectLiteR framework. *Methodology* 17, 307–325. doi: 10.5964/meth.7379

Kuiper, R. M., and Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychol. Methods* 15, 69–86. doi: 10.1037/a0018720

Mayer, A., and Dietzfelbinger, L. (2019). *EffectLiteR: Average and conditional effects*. R package version 0.4–4.

Mayer, A., Dietzfelbinger, L., Rosseel, Y., and Steyer, R. (2016). The EffectLiteR approach for analyzing average and conditional effects. *Multivariate Behav. Res.* 5, 374–391. doi: 10.1080/00273171.2016.1151334

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R foundation for statistical computing.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. New York, NY: Wiley.

SAS Institute Inc. (2015). *Sas/stat ® 14.1 User'S Guide*. Cary, NC: SAS Institute Inc..

Seber, G. A. F., and Lee, A. J. (2012). *Linear Regression Analysis*. Hoboken, NJ: Wiley.

Silvapulle, M. J., and Sen, P. K. (2005). *Constrained Statistical Inference: Order, Inequality, and Shape Restrictions*. Hoboken, NJ: Wiley.

Silvapulle, M. J., and Silvapulle, P. (1995). A Score test against one-sided alternatives. *J. Am. Stat. Assoc.* 90, 342–349. doi: 10.1080/01621459.1995.10476518

Turlach, B. A., and Weingessel, A. (2019). *Quadprog: Functions to solve quadratic programming problems*. R package version 1.5–8.

Vanbrabant, L. (2020). *Restriktor: Constrained statistical inference*. R package version 0.2–800.

Vanbrabant, L., Van de Schoot, R., and Rosseel, Y. (2015). Constrained statistical inference: Sample-size tables for ANOVA and regression. *Front. Psychol.* 5, 1565. doi: 10.3389/fpsyg.2014.01565

White, H. (1980). A heteroskedasticity-consistent covariance matrix and a direct test for heteroskedasticity. *Econometrica* 48, 817–838. doi: 10.2307/1912934

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *J. Stat. Softw.* 16, 1–16. doi: 10.18637/jss.v016.i09

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