Editorial

Confirmatory Factor Analyses in Psychological Test Adaptation and Development

A Nontechnical Discussion of the WLSMV Estimator

Kay Brauer[1], Jochen Ranger[1], and Matthias Ziegler[2]

[1]Department of Psychology, Martin Luther University Halle-Wittenberg, Halle, Germany
[2]Department of Psychology, Humboldt-Universität zu Berlin, Germany

The importance of providing structural validity evidence for test score(s) derived from psychometric test instruments is highlighted by several institutions; for example, the American Psychological Association (2014) demands that evidence for the validity of an instruments' internal structure and its underlying measurement model must be provided before it is applied in psychological assessment. The knowledge about the latent structure of data obtained with tests addressing the major question "What is/are the construct[s] being measured" by psychological tests under investigation (Ziegler, 2014, 2020). The study of structural validity is typically addressed with factor analyses when the test scores reflect continuous latent traits. As most submissions to Psychological Test Adaptation and Development (PTAD) deal with the adaptation and further development of existing measures, authors typically test a measurement model that is based on theoretical considerations and prior findings on original versions (or adaptations) of the test under investigation. Our literature review of PTAD's publications showed that more than 90% of the articles contain at least one confirmatory factor analysis (CFA).

As editor and reviewers of PTAD, we appreciate that authors are rigorous in providing evidence on the structural validity of their tests' data. However, since PTAD's inception in 2019, we experience that one comment is frequently communicated to authors during the review process, namely, the request to adjust the analytic approach in CFA from maximum likelihood (ML) estimation toward using the mean- and variance-adjusted weighted least squares (WLSMV; Muthén et al., 1997) estimator to account for the ordinal nature of the data that psychological instruments typically generate on the item level. In this editorial, we discuss the rationale behind choosing the WLSMV estimator when analyzing test adaptations and developments that are based on ordinal categorical data and concisely illustrate the problems associated with using the ML estimator (potentially in combination with robust tests of model fit) for such data.

A Short Recap of Basic Confirmatory Factor Analysis Principles

CFA aims at testing a predefined assumption about the structure of data (e.g., items). In test construction and evaluation, the measurement model of each test score is such an assumption about the items reflecting the trait in question (Ziegler & Hagemann, 2015). In particular, this contains a hypothesis about which manifest indicators (i.e., items) should be loaded by which latent factor(s). The relations between items and latent variables are one of the laws Cronbach and Meehl (1955) list as part of the nomological net. This net also includes relations between latent variables, which could also be tested using CFA. In short, CFA allows for testing whether an a priori assumed structure fits with the observed data. As mentioned, submissions to PTAD typically rely on prior evidence that provides assumptions regarding the dimensionality (i.e., the number of factors) and the item-factor assignment.

For illustration, one might imagine that we want to examine the measurement model (or factor model) of a translation of a self-report questionnaire that consists of 10 items. Let us assume that prior theoretical assumptions and empirical evidence from previous studies suggest that the 10 items reflect two latent factors. For example, prior evidence might suggest that Items 1–5 should be loaded by

Psychological Test Adaptation and Development (2023), 4, 4–12
https://doi.org/10.1027/2698-1866/a000034

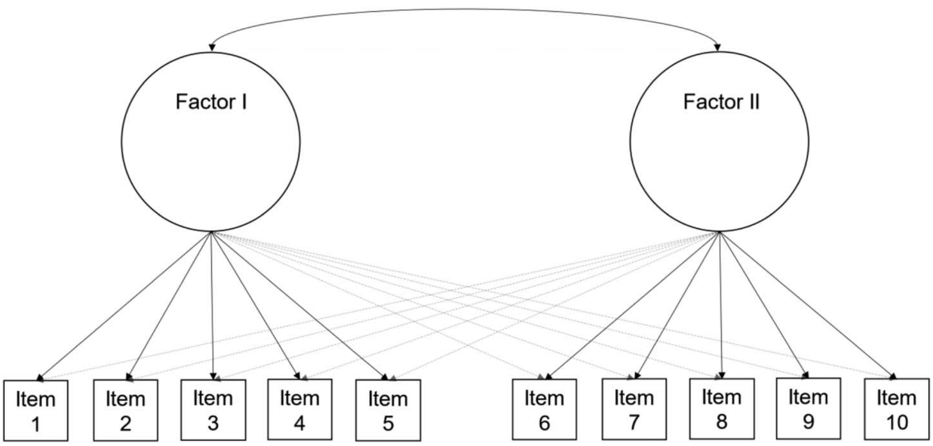

Figure 1. Confirmatory factor analysis model testing two correlated factors. *Note*. Thinned paths represent loadings that are set to zero. Residuals not displayed for simplicity. Double arrow indicates the interfactor correlation.

Factor I and Items 6–10 by Factor II. We specify the measurement model accordingly and as displayed in Figure 1, with loadings > 0 being allowed for Items 1–5 by Factor I and loadings > 0 for Items 6–10 by Factor II. At the same time, unintended loadings are restricted (i.e., loadings of Items 1–5 by Factor II and loadings of Items 6–10 by Factor I are set to be zero; thinned lines in Figure 1). When we assume reflective traits, the factors cause the intercorrelations among the set of items that are loaded by each respective factor. Thus, if the assumed model is correct, we would expect high intercorrelations among Items 1–5, as they share the same underlying factor, showing no substantial intercorrelations with Items 6–10, because the latter represents a distinct factor. On the other hand, we assume intercorrelations between Items 6 and 10 (Factor II), which are in turn not substantially correlated with Items 11–5.

CFA examines whether our hypothesized model and its implications for the covariance structure (the *implied* covariance structure among items) fit with the observed item responses (the *observed* covariance structure). In analyzing the data, we are interested in the magnitude and direction of the loading parameters and the overall model fit.

Model parameters are usually estimated with the maximum likelihood (ML) estimator. In ML estimation, one determines those parameter values (loadings, specific variances, and factor intercorrelations) that make the model implied covariance matrix as close as possible to the observed covariance matrix. As close as possible is defined via a discrepancy function that assesses the agreement between the observed and the model implied covariance matrix. The discrepancy function is zero when both matrices are identical and greater than zero otherwise. Thus, the smaller the discrepancy, the better does the assumed model explain the observed data structure. The discrepancy of ML estimation can be interpreted as the likelihood of the observed covariance matrix given the parameter values of the assumed model (Lawley & Maxwell, 1962). Hence, by minimizing the discrepancy, one determines those model parameter values that are most likely considering the observed covariance matrix and the specified model. Provided that the model holds, the ML estimator is consistent. The ML estimator is efficient for normally distributed data as in this case the sample covariances are sufficient statistics. Sufficient statistics are statistics that summarize the data without loss of information. More technically, the covariance matrix is the only aspect of the data that is relevant for the likelihood function, if the mean structure is ignored.

The degree of model-to-data fit is evaluated on the basis of goodness-of-fit indexes that inform about the absolute fit (e.g., the χ^2 value and root-mean-square error of approximation) and relative fit (i.e., relative to alternative models; e.g., Tucker–Lewis index and comparative fit index). These fit indexes allow gauging how well the assumed model can reproduce the observed variance–covariance matrix between the items. The fit indexes are typically evaluated on the basis of cutoff values (e.g., Hu & Bentler, 1999; see also Greiff & Heene, 2017; Heene et al., 2011; Hopwood & Donnellan, 2010, for a critical discussion). In case the fit indexes suggest good fit, we would conclude that the data reflect the assumed structure well, with the loadings exceeding zero substantially on their intended factor while being zero on their unintended factor.[1] On the contrary, if the model does not fit the data, we would, for example, test alternative models or examine modification indexes and revise the model accordingly, and then test the revised model in an independent sample to

[1] Note that we simplify the interpretation of the CFA here. One is typically not only interested in the fit index but also examines factor loadings, specificities, communalities, and factor intercorrelations.

avoid overfitting the model to the data of a single sample (Fokkema & Greiff, 2017).

While our description recapitulates the general proceeding in analyzing and interpreting the CFA, it must be noted that the accuracy of the estimations of parameters of the tested factor model and their *SE*s also rely on the choice of the estimation method. One guiding principle of choosing the best-suited estimator for CFA is to examine whether the assumptions of the estimator are met by the type of data analyzed.

Estimating the CFA Model on the Basis of Continuous and Discrete Data

Maximum Likelihood and Robust Maximum Likelihood Estimation

The ML estimator is provided as the default estimator in numerous standard statistical software packages. ML is consistent as long as the model and mild regularity conditions hold. These conditions include the following:

- The model is correctly specified, and the observed data are generated from this model.
- The observed variables are independent and have continuous distributions.
- The model parameters are independent and have continuous distributions.
- The sample size is large enough for the maximum likelihood estimator to be consistent.
- The model parameters are identifiable, meaning that they can be accurately estimated from the observed data.

The ML estimator provides accurate estimates in many scenarios dealing with continuously distributed variables (i.e., in multivariate normal distributed data). An example for an indicator that provides continuous response data is the visual analog scale, where test takers place their responses on a continuum between two poles (e.g., between 0% and 100%; Flynn et al., 2004). However, the ML estimator has limitations when analyzing data that do not follow a continuous distribution, such as ordered categorical data.

Categorical data are generated by items that are answered using ordered categories, such as in dichotomous indicators (e.g., items in ability tests are scored as 0 [*incorrect*] and 1 [*correct*]; e.g., Gnambs et al., 2021) and rating scales containing 3 to *k* response options, with anchors such as 1 (*strongly disagree*) and *k* (*strongly agree*; e.g., Dierickx et al., 2020). Many assessment instruments in psychology and neighboring sciences generate this type of discrete data (Simms et al., 2019). Likert (1932) proposed to use five response options, and although he did not provide an explanation from a psychometric point of view for this choice, numerous scales have adopted this suggestion (Simms et al., 2019). Accordingly, most psychological tests that are evaluated in submissions to PTAD are based on item responses that are dichotomous or of ordinal nature.

As noted, the ML estimator has several merits, as it is asymptotically unbiased, consistent, and efficient. However, these attributes only hold when certain assumptions are met as listed above. For example, the responses, and thus, the observed data to be analyzed, should follow a continuous and multivariate normal distribution (see, e.g., Bollen, 1989). These assumptions are violated when analyzing discrete data collected with rating scales that contain ordered categories with discrete and only few response options.

To illustrate the distribution of categorical responses, we want to examine the distributions of empirical data that were generated by $N = 540$ participants who responded to an item assessing extraversion. Each participant responded to the same item four times, namely, using a 2-, 4-, 6-, and 8-point rating scale with the end poles *does not apply* and *applies very much*, respectively.[2] The data provide us with an overview of how responses are typically distributed when using a standard response scale with frequently used response options (cf. Simms et al., 2019). Figure 2 shows the frequency distributions for the 2-, 4-, 6-, and 8-response option versions for the same item. The figure nicely portrays the discrete distributions resulting from the discrete response options. Of course, this is especially visible when checking the responses to the 2-point and 4-point rating scales (upper half of Figure 2), where we can see that the frequencies represent the realization of discrete events that indicate if a participant chose a response option *k*. Of course, the distribution is affected by the actual item difficulty, which empirically affects the distribution of the responses. An additional concern with discrete data from ordinal rating scales is that empirical realizations representing all response options are limited. For example, we see that response option 1 is rarely chosen, even when using multiple response categories: only 3.7% ($n = 20$) chose option 1 when responding to the 4-point rating scale, 1.3% ($n = 7$) in the 6-

[2] Item 6 of the German version of the *Big Five Inventory-Short* (BFI-S; Rammstedt & John, 2005). The data are taken from an ongoing study testing effects of the response format (Brauer et al., 2022).

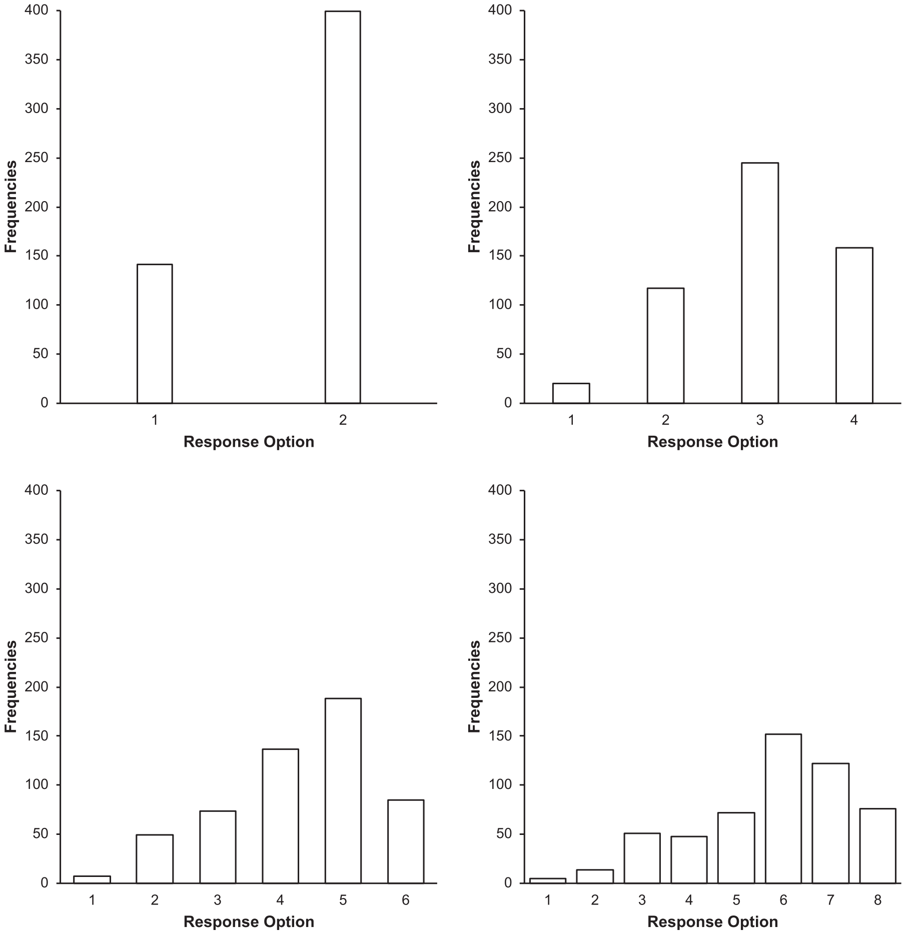

Figure 2. Histograms of responses to an extraversion item of N = 540 participants using 2-, 4-, 6-, and 8-point rating scales.

point format, and 0.9% (n = 5) in the 8-point format. Thus, some categories might be under-represented.

The categorical nature of the data goes along with the misspecification of the factor analytic model. In other words, the responses collected with rating scales that are popular and frequently used in the field are not *continuously* distributed but discrete. Hence, important assumptions required by the ML estimator are not satisfied when analyzing discrete data, and in consequence, the estimations of factor loadings, *SE*s, and model fit indexes are potentially biased (e.g., Beauducel & Herzberg, 2006; Kaplan, 2009; Li, 2016). However, if we base our interpretation of findings on a measurement model with inaccurate estimates, it increases the likelihood of making erroneous conclusions about the model. Thereby, our conclusions about the structural validity of the test under investigation can be erroneous, too.

The robust maximum likelihood (MLR) estimator has been introduced as an alternative. MLR eases the assumption of normality (Bollen, 1989). In brief, the MLR approach estimates model parameters with the regular ML approach but uses statistical corrections to the *SE*s and χ^2 model fit statistic (for details, see, e.g., Chou et al., 1991; Satorra & Bentler, 1994; Yuan & Bentler, 1998). While the MLR estimator eases the assumption of normality, it still requires data to be continuous, and thus, its suitability for ordered rating scales producing distinct data is still debatable. In a comprehensive simulation study, Bandalos (2014) concluded that robust ML might be considered a viable alternative but recommends another estimator.

WLSMV Estimation

The WLSMV estimator has been introduced to account for the ordinal nature of data as produced by ordered rating scales (Muthén et al., 1997). In WLSMV estimation, the ordinal response is interpreted as a result of a categorization process, which describes how test takers respond to an item: It is assumed that each response option on the observed level defines a range on a continuum of the response on the latent level. Whenever the latent response value is within a certain range, the corresponding response option on the observed level is chosen. We will illustrate this process on the basis of simulated data. Figure 3 contains simulated latent continuous responses to two

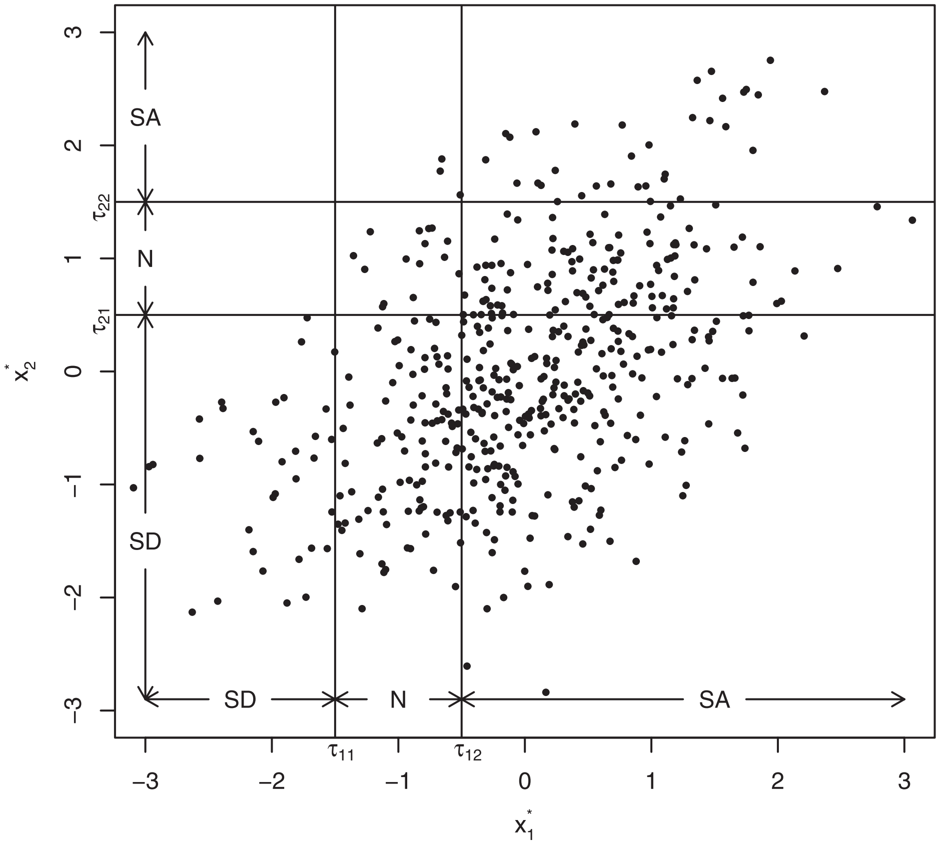

Figure 3. Continuous latent responses and categorized observed responses for two items. *Note*. Dots represent continuous responses for 500 participants in two items. τ_{11} to τ_{22} denote the thresholds that are employed when the continuous response is categorized into the three response options: SA = strongly agree, N = neutral, and SD = strongly disagree.

items (denoted as x_1 and x_2) by $N = 500$ test takers. The continuous responses might, for example, reflect the amount of agreement to the items of a personality questionnaire.

When responding to the items of the questionnaire, the test takers' latent continuous responses are mapped to the categories of the rating scale as follows. As displayed in Figure 3, the response labels define a range on the latent response continuum. In our example, we assume that the data are based on responses to three response categories (*strongly disagree*, *neutral*, and *strongly agree*). Whenever the latent continuous response falls in the range of one of the response labels, the response category is chosen. For example, if a test taker's response to item x_1 lies between the values of -1.5 and -0.5 on the latent continuum, this would represent choosing the *neutral* response option on the observed level. These ranges are defined by thresholds that are denoted by the Greek letter τ. There are $k - 1$ thresholds, defining the ranges between the k discrete response options. In Figure 3, the ranges are defined by the thresholds τ_{11} and τ_{12} for Item x_1, and τ_{21} and τ_{22} for item x_2. All test takers with a value below τ_{11} on the latent continuum choose the first discrete response option *strongly disagree* for item x_1. Those with a value between τ_{11} and τ_{12} choose response option *neutral*, and those with a value above τ_{12} choose the third response option (*strongly agree*). In the same way, responses to item x_2 are chosen on the basis of the location of the score on the latent continuum in relation to the thresholds τ_{21} and τ_{22}. However, note that the scatterplot displayed in Figure 3 cannot be observed directly. Instead, the categorized responses can be tabulated from the observed data. Table 1 gives the cross-tabulation of the observed categorized data from Figure 3.

Based on the simulated latent response data from Figure 3 and the observed discrete data provided in Table 1, we can estimate and compare the correlations between the two latent dimensions and the manifest answers to items x_1 and x_2. The correlation between the items when considering the observed categorical data is $r = .24$, whereas the correlation between the continuous responses is $r = .49$. Thus, the correlation of the observed categorical data is substantially lower than the correlation of the latent continuous data ($z = 4.59$, $p < .001$).

In general, the correlations (or covariances) between observed discrete data are not a valid estimate for the correlations between the continuous latent responses. This is problematic in case we use the observed discrete

Table 1. Cross-tabulation of frequencies of observed discrete responses to items x_1 and x_2

		x_1		
Item	Response	Strongly disagree	Neutral	Strongly agree
x_2	Strongly disagree	35	87	222
	Neutral	0	15	104
	Strongly agree	0	3	34

Note. N = 500 simulated responses.

data to estimate a factor model for the continuous latent responses. The observed covariance matrix does not converge to the true covariance matrix of the latent responses. The WLSMV estimator, however, aims at recovering the correlation between the continuous latent responses from the cross-tabulations of all items and their observed responses (see Muthén et al., 1997, for technical details). As the correlation and the thresholds determine the table frequencies, it is possible to estimate the correlations from the cross-tabulations and to account for the issue of dealing with noncontinuous data. Such correlations are denoted as tetrachoric correlations in case the discrete variables are binary or as polychoric correlations when the variables have more than two categories. The recovered correlation matrix is then used to determine the parameters of the factor model. In WLSMV estimation, one proceeds as in ML estimation by determining those parameter values that make the model implied correlation matrix as similar as possible to the recovered correlation matrix (which is based on the observed one). Similarity is again assessed by a discrepancy function. The discrepancy function of WLMSV estimation, however, differs from the function used in ML estimation to account for the fact that the data are not normally distributed. In its core, the discrepancy is assessed by the sum of the squared differences between the recovered and implied correlations. The squared differences are weighted to increase the efficiency of the estimator. The WLMSV is consistent and asymptotically normally distributed (Muthén et al, 1997). It can be complemented by corresponding tests of model fit that parallel the ones used in robust ML estimation.

In consequence, estimates for the model parameters depend on the correlations among items and the estimator used. In other words, we can infer that the estimates for the CFA are affected by whether the same data are continuous or categorized into discrete options. With regard to distributional assumptions, WLSMV assumes that the continuous latent responses follow a multivariate normal distribution, which is in line with assumptions for the majority of constructs studied in psychology (Li, 2016), whereas the observed data are not required to be normally distributed (Muthén et al., 1997). Thus, the WLSMV estimator is specifically designed to analyze the ordinal response data mostly generated by psychological measures.

Comparisons of the Performance of ML, MLR, and WLSMV

The choice of the estimator is important for the accurate estimation and interpretation of CFAs that are used to draw conclusions about structural validity. We aimed at providing a nontechnical understanding of the issues when applying the ML and MLR estimators to discrete data in CFA and discussed the WLSMV estimator that has been specifically introduced to deal with categorical data.

Simulation studies have addressed the accuracy of factor loadings, interfactor correlations, their respective standard errors, convergence rates, and fit indexes in relation to sample size, model complexity, in normal and non-normal distributions, and with regard to the number of response categories for the ML, MLR, and WLSMV estimators. We want to shortly discuss two important studies from the field that compared the estimators focused here.

Beauducel and Herzberg (2006) examined the performance of the ML and WLSMV estimator for indicators that were answered with 2, 3, 4, 5, and 6 response categories in samples containing responses by N = 250, 500, 750, and 1,000 simulated respondents. They fixed the loading to .50 (oblique models) and .55 (orthogonal models) and found that the WLSMV outperformed the ML estimator in case of few response categories (i.e., 2 and 3 response options) with regard to more accurate estimates of loadings, standard errors, and fit indexes. Moreover, Beauducel and Herzberg found that loadings were estimated more accurately by the WLSMV estimator, that is, ML underestimated loadings and yielded higher standard errors in comparison to the WLSMV estimator when analyzing categorical data, irrespective of the number of response options. While their findings provided initial systematic evidence on the performance of ML and WLSMV estimators, it must be noted that their simulations did only consider approximatively normally distributed responses and did not consider their performance when distributions are non-normal (i.e., characterized by robust skewness and/or kurtosis), which is of particular importance when analyzing data generated from ordered categories.

To our knowledge, Li (2016) provided the most comprehensive simulation study to date, comparing the most frequently used estimators for ordered noncontinuous data (MLR and WLSMV). They examined the performance of the estimators in sample sizes of N = 200, 500, and

1,000 participants, slightly and moderately non-normal distributed data sets, and concerning whether 4, 6, 8, or 10 response categories were used. In short, Li concluded that WLSMV outperformed MLR concerning the estimation of factor loadings (MLR underestimated loadings), irrespective of the number of response categories, sample sizes, and distribution characteristics. Factor intercorrelations were slightly overestimated by WLSMV when sample size is small (i.e., ≤ 200) and/or the underlying latent distribution deviates from normality in comparison to MLR. Standard errors of factor loadings were sensitive to sample size in MLR and WLSMV alike, but MLR outperformed WLSMV when sample size is small ($N = 200$), but differences were negligible in larger samples. Finally, the χ^2 fit index tends to over-reject the models for both WLSMV and MLR. Taking the findings together, Li concludes that findings derived with MLR should be interpreted cautiously due to the bias in parameter estimates in combination with small standard errors. While Li extended the knowledge on the performance of WLSMV and MLR on the basis of ordered categorical responses that follow normal and non-normal distributions, their simulations are also not free from restrictions that also limit the interpretations. Most importantly, it must be noted that the simulations assumed loadings of .70 for each item, which is uncharacteristically high for rating scale items.

When combing the knowledge from those studies, the findings from Beauducel and Herzberg (2006) and Li (2016) converge regarding the WLSMV estimator yielding more accurate loadings than the ML-based estimators.

Recommendation and Conclusion

MLR and WLSMV estimators have their advantages and disadvantages. Thus, there is no one-solution-fits-all recommendation, and it depends on the data and what researchers are interested in when analyzing structural validity with CFA. From our experience, the majority of submissions to PTAD are interested in evaluating the loading structure of their assumed models. Considering the current knowledge in the field (e.g., Beauducel & Herzberg, 2006; Li, 2016), authors might want to favor the WLSMV estimator when the major focus lies on accurately estimating the factor loadings on the basis of categorical data. On the contrary, MLR provides more accurate estimates of factor interrelations than WLSMV in many cases. Thus, if a study's main aim is in testing hypotheses concerning the factor interrelations instead of primarily testing factor loadings, the MLR estimator might be favored.

One could argue that it could be fruitful to analyze data generated by responses to ordered categorical rating scales with both approaches MLR and WLSMV and *transparently* reporting findings and their convergence across methods. The findings of ML(R) and WLSMV estimations should overlap comparatively well when the analyzed data follow a normal distribution (e.g., Beauducel & Herzberg, 2006). On the contrary, if findings from WLSMV and ML analyses do not converge, Li's (2016) study could provide us with hints and guide authors in checking whether certain features of their data might be responsible for differences across estimation methods. The latter could contribute to expand the knowledge and inform future research of the factorial validity of a test under investigation. In general, we suggest contextualizing the findings of factor analyses of a test adaptation in relation to prior findings such as the original test and alternative adaptations.

Finally, it must be noted that CFA is only one potential approach to investigate the structural validity of measures that generate noncontinuous data. Alternatively, the field of item response theory (IRT) offers approaches to examine the trait structure underlying tests as well (Bond et al., 2020; van der Linden, 2016). In fact, it has been shown that the assumptions made by the WLSMV estimator and the graded response model are similar (Takane & de Leeuw, 1987). Item response models, however, are ideally estimated with full information approaches, which are more efficient than the WLMSV estimator (Forero & Maydeu-Olivares, 2009). However, one must consider that the complexity of IRT models often requires substantially increased computational power and time in comparison to CFA, when testing high-dimensional models. The computational time increases nonlinearly also depending on the estimator and the dimensionality of the test. However, we encourage that authors submitting to PTAD consider IRT analyses as one alternative to the classical CFA framework.

In conclusion, our discussion is a response to the observation that PTAD receives many submissions that rely on the nonrobust ML estimator when analyzing ordered categorical data. We can only speculate, but one reason for this observation might be that popular software packages such as *Mplus* (Muthén & Muthén, 1997–2017), AMOS (Arbuckle, 2019), and CRAN R's *lavaan* (Rosseel, 2012) use the ML estimator as preset default estimation method which can be used conveniently. However, there is no certainty that the ML estimator provides accurate findings when treating responses as approximately continuous and ignoring the categorical and noncontinuous nature of the data. Considering classical and recent findings (e.g., Beauducel & Herzberg, 2006; Li, 2016), WLSMV seems to provide more accurate estimates of loadings when data are generated through categorical responses. We hope that our discussion contributes to provide an understanding of why editors and reviewers working for PTAD request authors to consider using the WLSMV estimator for their CFAs.

References

American Psychological Association. (2014). *Standards for educational and psychological testing*. AERA Publications.

Arbuckle, J. L. (2019). *Amos* (*Version 26.0*) [Computer program]. IBM SPSS.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling*, *21*(1), 102–116. https://doi.org/10.1080/10705511.2014.859510

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186–203. https://doi.org/10.1207/s15328007sem1302_2

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Bond, T., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Brauer, K., Nussbeck, F. J., Zwiky, E., & Proyer, R. T. (2022). *Testing effects of the response scale on indicators of interpersonal perception* [Manuscript in preparation].

Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 347–357. https://doi.org/10.1111/j.2044-8317.1991.tb00966.x

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Dierickx, S., Smits, D., Corr, P. J., Hasking, P., & Claes, L. (2020). The psychometric properties of a brief Dutch version of the Reinforcement Sensitivity Theory of Personality Questionnaire. *Psychological Test Adaptation and Development*, *1*, 20–30. https://doi.org/10.1027/2698-1866/a000004

Flynn, D., van Schaik, P., & van Wersch, A. (2004). A Comparison of multi-item Likert and Visual Analogue Scales for the assessment of transactionally defined coping function. *European Journal of Psychological Assessment*, *20*(1), 49–58. https://doi.org/10.1027/1015-5759.20.1.49

Fokkema, M., & Greiff, S. (2017). How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it. *European Journal of Psychological Assessment*, *33*(6), 399–402. https://doi.org/10.1027/1015-5759/a000460

Forero, C., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275–299. https://doi.org/10.1037/a0015825

Gnambs, T., Scharl, A., & Rohm, T. (2021). Comparing perceptual speed between educational contexts: The case of students with special educational needs. *Psychological Test Adaptation and Development*, *2*, 93–101. https://doi.org/10.1027/2698-1866/a000013

Greiff, S., & Heene, M. (2017). Why psychological assessment needs to start worrying about model fit. *European Journal of Psychological Assessment*, *33*(5), 313–317. https://doi.org/10.1027/1015-5759/a000450

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*(3), 332–346. https://doi.org/10.1177/1088868310361240

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kaplan, D. (2009). *Structural equation modeling: Foundations and extensions* (2nd ed.). Sage.

Lawley, D., & Maxwell, A. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society: Series D*, *12*(3), 209–229. https://doi.org/10.2307/2986915

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 44–53.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Retrieved from https://www.statmodel.com/download/Article_075.pdf

Muthén, L. K., & Muthén, B. O. (1997–2017). *Mplus user's guide*. Muthén & Muthén.

Rammstedt, B., & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit [Short version of the Big Five Inventory (BFI-K): Development and validation of an economic inventory for assessment of the five factors of personality]. *Diagnostica*, *51*(4), 195–206. https://doi.org/10.1026/0012-1924.51.4.195

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye, & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, *31*(4), 557–566. https://doi.org/10.1037/pas0000648

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/BF02294363

van der Linden, W. J. (2016). *Handbook of item response theory*. Taylor & Francis. https://doi.org/10.1201/9781315119144

Yuan, K.-H., & Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 289–309. https://doi.org/10.1111/j.2044-8317.1998.tb00682.x

Ziegler, M. (2014). Stop and state your intentions! Let's not forget the ABC of test construction. *European Journal of Psychological Assessment*, *30*(4), 239–242. https://doi.org/10.1027/1015-5759/a000228

Ziegler, M. (2020). Psychological test adaptation and develop ment – how papers are structured and why. *Psychological Test Adaptation and Development*, *1*, 3–11. https://doi.org/10.1027/2698-1866/a000002

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items: Pitfalls and loopholes. *European Journal of Psychological Assessment*, *31*(4), 231–237. https://doi.org/10.1027/1015-5759/a000309

Published online February 7, 2023

Acknowledgments
We are grateful to Rebekka Sendatzki for her help in preparing the figures.

ORCID

Kay Brauer
https://orcid.org/0000-0002-7398-8457
Jochen Ranger
https://orcid.org/0000-0001-5110-1213
Matthias Ziegler
https://orcid.org/0000-0003-4994-9519

Kay Brauer
Department of Psychology
Martin Luther University Halle-Wittenberg
Emil-Abderhalden-Str. 26-27
06099 Halle
Germany
kay.brauer@psych.uni-halle.de