Manuscript accepted for publication as a commentary to:
Hoogeveen, S. et al. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*.

Why researchers should not ignore measurement error and skewness in questionnaire item scores

Paul Lodder[1,2]

[1] Department of Methodology and Statistics, Tilburg University

[2] Department of Medical and Clinical Psychology, Tilburg University

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Researchers commonly study associations between latent variables measured with items showing ordinal and skewed score distributions. For instance, in the many analysts religion project (The MARP Team, 2022) such distributions apply to the item scores measuring the latent variables religiosity, perceived cultural norms and well-being (see supplementary Figure 1). Researchers can estimate associations between latent variables according to several statistical methods that vary for instance in how they handle the presence of skewness and measurement error in the item scores (Lodder et al., 2019). Choices made by researchers during the analysis process can influence the conclusions drawn from statistical analyses (Wicherts et al., 2016). Indeed, differences between MARP teams in such analysis choices explains the heterogeneity in estimated effect sizes.

When researchers operationalize latent variables as the mean or sum of item scores then they assume that these latent variables are measured without error. Including such observed scores in regression analyses generally produces underestimated latent variable associations, a phenomenon called attenuation bias (Spearman, 1904). This bias is especially relevant when testing interaction effects because the product term typically used to model an interaction effect not only multiplies true score variance, but also measurement error variance, thus further reducing the reliability of the interaction term (Busemeyer & Jones, 1983). Besides containing measurement error, the item scores of psychological questionnaires are often not normally distributed (e.g., Reise & Waller, 2009). Using statistical models that incorrectly assume such skewed ordinal items to be continuous and normally distributed can result in biased parameter estimates (Dolan, 1994; Rhemtulla, Brosseau-Liard, & Savalei, 2012). In this commentary I use a computer simulation to illustrate that ignoring skewness or measurement error in questionnaire item scores often results in biased effect estimates, especially when testing interaction effects. These simulation findings may explain some heterogeneity in the effects estimated by the participating MARP teams.

I simulated 800 datasets with scores on two independent (X and Z) and one dependent (Y) latent variable. In the structural model I simulated either a main effect (X) or an interaction effect (X*Z) on Y. When generating item scores from a factor model, I varied item skewness by using symmetrical or asymmetrical threshold parameters that map the ordinal item responses to an underlying continuous latent item response (Flora & Curran, 2004). This resulted in four skewness conditions: (1) no skewed item scores; (2) positively skewed item scores for X and Z; (3) negatively skewed item scores for Y; (4) positively skewed item scores for X and Z, and negatively skewed item scores for Y.

Each of the 800 simulated datasets was analyzed using three methods: (1) a regression using the sums of observed item scores, (2) a structural equation model (SEM) assuming continuous item scores, and (3) a SEM for ordered categorical items (CATSEM), which is the approach that I have used to analyze the MARP data (Team 26: https://osf.io/m7hck/). Both SEM and CATSEM can handle the measurement error in the item scores when estimating the association between latent variables. SEM assumes that the item scores are continuous and linearly related to the measured latent construct, an assumption likely violated when modeling skewed ordinal data (Flora, LaBrish & Chalmers, 2012). CATSEM does not make this assumption because it nonlinearly maps the observed ordinal item scores to a continuous normally distributed latent response variable using a set of threshold parameters. For each simulation condition, I visualized the absolute- and relative bias using boxplots. Next, empirical MARP estimates were aggregated and compared across the three methods using a random effects meta-analysis. The R-script and a more detailed methods section can be found in the supplement.

Figure 1 shows for each method the bias in the estimated regression coefficients of the main- and interaction effects. The dashed grey lines indicate the acceptable margin of 10% relative bias. The results show that CATSEM on average produced acceptable estimates in all simulation conditions. SEM was unbiased when item scores were normally distributed, but it underestimated the true effect when the item scores of either the independent or dependent latent variables were skewed. Observed score regression produced underestimated effects, especially interactions. This bias increased when the item scores of independent variables were positively skewed.

To relate the simulation results to the estimates by the MARP teams, Figure 2 shows for each method the estimated standardized regression coefficients for the main effect of religiosity (left panel) and the interaction between religiosity and perceived cultural norms (right panel). The figure also indicates the overall effect size and 95% confidence interval estimated in a random effects meta-analysis including all teams using each method. There was considerable heterogeneity in the estimates of the observed score regression approaches. Nevertheless, on average this method produced significantly smaller interaction effects than the two SEM methods (B = -0.028, 95%CI = -0.051 to -0.005), while this difference for the main effects did not reach significance (B = -0.021, 95%CI = -0.059 to 0.017). The difference between the two SEM methods was not statistically significant, yet this comparison was underpowered due to the small number of teams using those methods.

The current findings illustrate that ignoring measurement error produces underestimated regression estimates, especially with respect to interaction effects. This result echoes earlier work showing that SEM produces less biased interaction effect estimates than observed score regression under a wide range of simulation conditions (Lodder et al., 2019). The current findings also suggest that ignoring skewness further increases the negative bias in the estimated main- and interaction effects. This resonates with previous work showing that SEM and observed score regression produce biased interaction effect estimates when item scores are ordinal and positively skewed (Lodder, Emons, Denollet & Wicherts, 2021). The current simulation findings may in part explain the differences between participating MARP teams in the estimated main- and interaction effects. The non-significant interaction effects reported by MARP teams that have used a linear regression on observed scores may in fact be false negative conclusions due to underestimated interaction effects. The current study highlights the importance of inspecting item score distributions and questionnaire reliability before choosing a statistical model. Researchers are recommended to use CATSEM when testing associations between latent variables that are measured with skewed ordinal item scores. Not doing so risks underestimated associations, especially when estimating interaction effects.

References

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93(3),* 549. doi:10.1037/0033-2909.93.3.549

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47(2),* 309–326. doi:10.1111/j.2044-8317.1994.tb01039.x

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods, 9(4)*, 466.

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology, 3,* 55.

Lodder, P., Emons, W. H., Denollet, J., & Wicherts, J. M. (2021). Latent logistic interaction modeling: A simulation and empirical illustration of Type D personality. *Structural Equation Modeling: A Multidisciplinary Journal, 28(3),* 440-462.

Lodder, P., Denollet, J., Emons, W. H., Nefs, G., Pouwer, F., Speight, J., & Wicherts, J. M. (2019). Modeling interactions between latent variables in research on Type D personality: A Monte Carlo simulation and clinical study of depression and anxiety. *Multivariate behavioral research, 54(5),* 637-665.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, 38(2), 171–189. doi:10.1111/j.2044-8317.1985.tb00832.x

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology, 5(1),* 27–48. doi:10.1146/annurev.clinpsy.032408.153553

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17(3),* 354. doi:10.1037/a0029315

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15(1),* 72–101. doi:10.2307/1412159

The MARP Team (2022). A Many-Analysts Approach to the Relation Between Religiosity and Well-being. Manuscript in preparation.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology, 7*, 1832.

Manuscript accepted for publication as a commentary to:
Hoogeveen, S. et al. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*.

***Figure 1**: For three methods (sum score regression vs. SEM vs. CATSEM), boxplots summarizing the estimated regression coefficients (upper row) and relative bias (bottom row) in the main effect of X and the interaction effect between X and Z, in four simulation conditions that varied in whether skewness was present in the item scores of X&Z and/or Y. In the upper row, the solid grey line indicates the true simulated value of the main or interaction effect. In the bottom row, the dashed grey lines indicate the 10% margin of acceptable relative bias.*

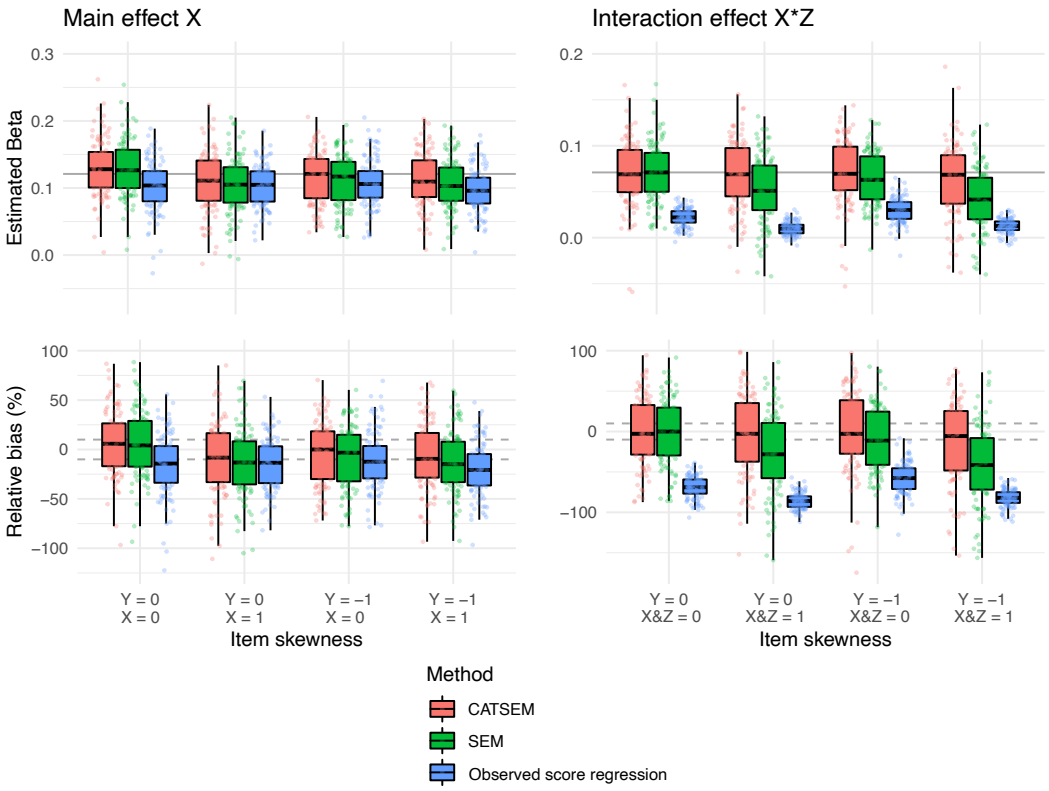

***Figure 2**: For three methods (observed score regression vs. SEM vs. CATSEM), the estimated standardized regression coefficients for the main effect of religiosity on well-being (left panel) and the interaction effect between religiosity and perceived cultural norms (right panel). Circle size corresponds to the standard error expressing the uncertainty in each estimate. Error bars indicate the 95% confidence interval of the overall effect estimate in a random effects meta-analysis for all effects estimated according to each method.*