# The Impact of Different Methods to Correct for Response Styles on the External Validity of Self-Reports

Anna Scharl & Timo Gnambs

Leibniz Institute for Educational Trajectories, Germany

## Author Note

Anna Scharl https://orcid.org/0000-0003-0081-1893

Timo Gnambs https://orcid.org/0000-0002-6984-1276

This paper uses data from the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands).

Correspondence concerning this article should be addressed to Anna Scharl, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany, E-mail: kontakt@amc-scharl.de.

## Abstract

Response styles (RSs) such as acquiescence represent systematic respondent behaviors in self-report questionnaires beyond the actual item content. They distort trait estimates and contribute to measurement bias in questionnaire-based research. Although various approaches were proposed to correct for the influence of RSs, little is known about their relative performance. Because different correction methods formalize the latent traits differently, it is unclear how model choice affects the external validity of the corrected measures. Therefore, the present study on *N* = 1,000 Dutch respondents investigated the impact of correcting responses to measures of self-esteem and need for cognition using structural equation models with structured residuals, multidimensional generalized partial credit models, and multinomial processing trees. The study considered three RSs: extreme, midpoint and acquiescence RS. The results showed homogeneous correlation patterns among the modeled latent variables and with external variables, especially, if those were not themselves subject to RSs. In that case, the IRT-based models, including an uncorrected model, still yielded consistent results. Nevertheless, the strength of the effect sizes showed variation.

*Keywords*: response styles, generalized partial credit model, processing trees, self-esteem, need for cognition

# The Impact of Different Methods to Correct for Response Styles on the External Validity of Self-Reports

The validity and reliability of personality instruments is influenced by numerous factors such as their content scope or specific item wordings (Soto & John, 2019). For example, wording effects have been shown to lead to response styles (RSs), that is, systematic behavioral tendencies determining self-report responses independently of the item content and the intended latent trait (Baumgartner & Steenkamp, 2001; Soto & John, 2019). Acquiescence (ARS) is the tendency to agree with an item regardless of its content (Plieninger & Heck, 2018), while the extreme RS (ERS) favors choosing the most extreme response categories and the midpoint RS (MRS) reflects a preference for the middlemost category (Falk & Cai, 2016). The consequences of RSs comprise distorted means and variances and, consequently, biased covariance structures and inferences (cf. van Vaerenbergh and Thomas, 2013).

Although the origins and effects of RSs (e.g., Baumgartner & Steenkamp, 2001) and how to statistically correct for them (e.g., Böckenholt & Meiser, 2017) have been investigated thoroughly, it may be difficult for applied researchers to choose between different approaches to deal with RSs. Previous comparative studies (e.g., Wetzel et al., 2016; Zhang, & Wang, 2020) examined only one or two RSs, but did not provide a comprehensive evaluation of more RSs. Therefore, this study will compare three methods correcting for ARS, MRS, and ERS. Since capturing the covariance structure of the content traits with third variables is paramount for psychological research, the focus of this study will be on the external validity of the estimated traits.

## Theoretical Background

### Response Styles and Response Style Correction Methods

RSs are prevalent in self-reported data and account for up to 29% of the variance in observed responses (Baumgartner & Steenkamp, 2001). RSs occur for different response formats (e.g., agree-disagree scales; Liu et al., 2015). Even forced-choice items that, by definition, cannot display RSs, show substantial correlations with RS indicators (He et al., 2014). Furthermore, different item characteristics can influence an item's RS propensity (Soto & John, 2019; Van Vaerenberg & Thomas, 2013). For example, rating scales with more response options were associated with increased MRS and ARS, but comparable or even decreased ERS (e.g., Kieruj & Moors, 2011; Kutscher & Eid, 2020). Additionally, RSs exhibit moderate stability within persons over time (Liu et al., 2015; Weijters et al., 2010) and were found to be systematically associated with diverse respondent characteristics (Van Vaerenbergh & Thomas, 2013). Thus, it has been argued that RSs constitute trait-like constructs. Therefore, various statistical post-hoc methods were developed to control for RSs in empirical data if RSs cannot be eliminated during item construction (Soto & John, 2019; Van Vaerenberg & Thomas, 2013).

An early approach uses manifest indicators to correct for RSs (Greenleaf, 1992). If the questionnaire contains items beyond the target scale, RS indicators can be calculated from them (Baumgartner & Steenkamp, 2001). These indicators are derived by recoding the items according to their RS pattern (e.g., for MRS, an item is coded as 1 if the middle category was used and 0 otherwise). The means of the RS indicators constitute the observed RS indices. They can be added as additional variables to factor analyses (He et al., 2014) or used as regressors for the content items in structural equation models (SEMs). This method is very versatile and can incorporate as many RSs as theoretically necessary. However, it does not estimate RS traits and, therefore, does not allow conclusions about the nomological net of RSs. Also, additional items must be included in the survey to calculate these indicators.

Methods based on item response theory (IRT), however, can simultaneously estimate content and RS traits based on the content questionnaires alone (Wetzel & Carstensen, 2017). Both multidimensional IRT (MIRT) models and item response trees (IRTrees; Plieninger, 2021) have been used to correct for several RSs. For example, Wetzel and Carstensen (2017) applied a multidimensional partial credit model (PCM) to model a content trait, ERS and MRS. Falk and Cai (2016) extended this approach to model ERS, MRS and ARS with a multidimensional nominal response model (see Figure S1 in the supplement for an illustration of the model). The common idea underlying these IRT models for categorical responses is the estimation of item and participant characteristics that determine the probability of endorsing an item category (e.g., 3 in a rating scale from 1 to 7; Reckase, 2009). RSs are modeled as person characteristics next to the content trait. They are distinguished from one another by a scoring function similar to the RS indicators' (Falk & Cai, 2016; see the supplement for more information). RSs can also be thought of as a mean shift in the threshold parameters of MIRT models, leading to different solution probabilities for the same trait values (Henninger & Meiser, 2020).

In contrast, IRTrees regard item responses as the result of a decision process. One example process could be as follows: The respondents first decide for or against using the middle category (i.e., MRS) and then whether to endorse the item or not (i.e., ARS). If ARS is present, the decision for or against the extreme category follows (i.e., ERS). Otherwise, the respondents' trait value leads to a decision to endorse the item or not, which is then, again, governed by ERS. These decision processes can be represented by tree structures (Böckenholt & Meiser, 2017). IRTrees model the likelihood of following a tree branch to its leaf (i.e., choosing a category) by modeling the probability for a decision using IRT. If multiple branches lead to the same category, the branch probabilities are summed up in a so-called multinomial processing tree (MPT; Plieninger & Heck, 2018). The MPT described above is depicted in Figure S2. Here ERS as a response tendency is represented by ordinal decision nodes (Meiser, Plieninger, & Henninger, 2019). Based on the tree diagrams, the models can be expressed in a series of equations that describe the probabilities of deciding on an item category (Figure S2B). These decision probabilities can be described by IRT models, assigning each decision an RS propensity for both test takers and items (Plieninger, & Heck, 2018). In this study, we modeled ERS following ARS and ERS following the trait-based decision as one. Therefore, a person's tendency to choose more extreme values is not dependent on their tendency to agree with an item, but is of a general nature (cf. Plieninger & Heck, 2018). Similarly, an item's ERS propensity is modeled as independent of ARS.

## Comparison of Correction Methods

IRT-based models are as versatile as manifest indicator models. Additionally, they offer insights into the variance-covariance structure of content and RS traits as well as the relation of RSs to other constructs. Moreover, IRT models lower the burden on participants by not requiring additional items to form RS indicators. However, these models require larger sample sizes.

Although different correction methods have been compared previously, most studies were limited to selected RSs. For example, Zhang and Wang (2020) evaluated ERS and MRS, while the studies by Primi et al. (2019) and Wetzel et al. (2016) only focused on ARS or ERS. In these studies, MIRT methods performed at least as well as the other methods. Still, little is known regarding the effect of RS corrections on validity correlations. While Zhang and Wang (2020) suggested that RS corrections for ERS and MRS (compared to no correction) did not affect the correlation between a content trait and an external criterion, Primi et al. (2020) found that failing to properly correct for ARS severely impacted criterion validities. Thus, existing findings on the impact of RS corrections in self-report instruments on validity correlations are mixed and limited to few RSs.

## Present Study

The present study extends previous work on the impact of RS correction methods on the validity of the measured trait. In contrast to prior research, we acknowledge ERS, MRS, and ARS and simultaneously correct for them using three popular correction methods (i.e., RS indicators, MIRT, and MPTs). We compare the variance-covariance structure of content traits for the need for cognition scale (NFCS; Cacioppo et al., 1996) and Rosenberg's (1965) self-esteem scale (RSES) with demographic variables as well as personality traits across different corrected and uncorrected models. The choice of external variables examined in the current study was guided by previous validity research for these domains which found small to medium correlations between age and NFCS ($r$ = -.45 to $r$ = .10; Cacioppo et al., 1996) or RSES ($r$ = .15; Franck et al., 2008), and negligible correlations for gender (Cacioppo et al., 1996; Pullmann & Allik, 2000; Sinclair et al., 2010). Moreover, longer education was found to be positively associated with both scales (e.g., Cacioppo et al., 1996; Sinclair et al., 2010). Construct validity with regard to the Big Five showed that the NFCS correlated most strongly with neuroticism and openness to experience and to a lesser degree also extraversion (Cacioppo et al., 1996), while the RSES showed more positive correlations with extraversion, openness, and conscientiousness ($r$ = [-.12; .47]) and mixed directions for the other facets ($r$ = [-.69; .69]; e.g., Franck et al., 2008; Pullmann & Allik, 2002; Schmitt & Allik, 2005). Finally, Cacioppo and colleagues (1996) also reported that NFCS and RSES correlated between $r$ = .15 to .42 with each other. Therefore, the present study addresses two research questions using a Dutch population sample. First, how are the correlations between the latent content traits and the RS traits affected by the correction method? Second, how are the correlations between the latent content traits and validity criteria affected by the correction method?

# Methods

## Participants

The participants were part of the *Longitudinal Internet Studies for the Social Sciences* panel (LISS; Scherpenzeel & Das, 2010), a representative sample of Dutch individuals. We focused on a random subsample of $N$ = 1,000 respondents participating in the first wave in 2008 who provided a complete set of answers on the personality instruments. Their mean age was $M$ = 45.97 years ($SD$ = 15.73), ranging from 16 to 90 years. More than half (55%) were female. About 10% did not have a traditional educational degree, about 37% finished their school education and around 52% had completed vocational or academic training.

## Material

**Target constructs.** The RSES consisted of 10 items rated on 7-point Likert scales from 1 (strongly disagree) to 7 (strongly agree). Because of low cell frequencies for negative extreme categories, we collapsed the categories 1 and 2 as well 6 and 7, resulting in 5-point item scores, which resulted in a less skewed response distribution (cf. Figure S9). The RSES scale mean was 3.35 with a standard deviation of 0.71. Cronbach's α was .87 and McDonald's $\omega_h$ was .91. The NFCS consisted of 18 items rated on 7-point Likert scales from 1 (strongly disagree) to 7 (strongly agree). The scale mean was 3.31 with a standard deviation of 0.91. Cronbach's α was .88 and McDonald's $\omega_h$ was .90.

**RS indicators.** Thirty items that were not part of the RSES or NFCS were randomly sampled from the LISS panel personality data. A subset of sufficiently unrelated ($|r| \leq .3$) items with similar proportions of RSs was formed for each RS. Then, indices for ARS, ERS, and MRS were calculated for each participant (Weijters et al., 2010). A detailed account of how the indicators were calculated is given in the supplement.

**External variables.** The participants' age was measured in years. Gender was recoded to 0 (males) and 1 (females). The educational level of the participants was recoded to "no traditional education" (0), "finished school education" (1), and "higher education" (2).

The Big Five were measured with 50 items from the International Personality Item Pool (Goldberg, 1999) on 5-point Likert scales from 1 (very inaccurate) to 5 (very accurate). Scale scores were created by averaging the item scores, after recoding negatively keyed items. The internal consistency was α = .76 and $\omega_h$ = .83 for openness, α = .77 and $\omega_h$ = .81 for conscientiousness, α = .86 and $\omega_h$ = .89 for extraversion, α = .80 and $\omega_h$ = .84 for agreeableness, and α = .88 and $\omega_h$ = .90 for emotional stability.

## Statistical Analyses

First, the RSES and NFCS were scaled with generalized PCMs (GPCM; Muraki, 1992) as a baseline. Then we corrected for ARS, ERS, and MRS using SEMs regressing the items on RS indices (Greenleaf, 1992), multidimensional GPCMs (MGPCM; Reckase, 2009), and MPTs (Plieninger & Heck, 2018). For the 5-category RSES, the model shown in Figure S2 simplifies to Plieninger and Heck’s (2018) model which was further restricted by equality constraints for the item parameters of ERS conditional on ARS and ERS conditional on the target trait to ensure model convergence in the RSES data. The restricted model could recover the response distribution adequately. The SEM’s metric was set by fixing the latent variable’s variance to 1. The MPT was identified by constraining the latent variable means to 0; in the (M)GPCM, the latent variable variances were additionally set to 1. Furthermore, all analyses but the MPT which used the original data were conducted using recoded data with appropriate scoring for the RS traits in case of the MGPCM. Next, latent trait scores were calculated for each participant as a set of 20 plausible values to approximate latent correlations between the target constructs and the external criterion variables. The research questions were addressed by computing correlation (for age, gender, the Big Five, RSES, and NFCS) and regression coefficients (for education, which was dummy coded with “no traditional education” as the reference category). Pairwise comparisons of the correlation and regression coefficients were carried out with Bonferroni corrected type I error rates to ensure a confidence level of 95%. A formal description of the models, MPT prior distributions, model fit diagnostics, and pairwise comparisons are given in the supplement.

# Results

## Reliability of Latent Traits

The modeling choice had a pronounced impact on the reliabilities of the latent traits, more so for the RSES than the NFCS (see Table 1). For the RSES, the content trait exhibited good reliabilities around .80 in the GPCM and SEM, while they were smaller for the MPT and MGPCM. The RSs generally exhibited better reliabilities for the MPT, while they fell as low as .30 for the MGPCM. The reliabilities of the NFCS were generally higher, but the RS traits under both MGPCM and MPT also exhibited sometimes weak reliability as low as .56. Detailed information on how the reliabilities were calculated is given in the online supplement.

## Correlations between Content and RS Traits

These correlations are only available for the IRT models that model latent RS traits (Table 2). For the NFCS under the MGPCM, the content trait was significantly correlated with MRS (*r* = -.107, *CI* = [-.168, -.045]) and ARS (*r* = .080, *CI* = [.018, .141]), and showed small positive correlations with ERS (*r* = .050, *CI* = [-.012, .112]). Under the MPT, the NFCS showed a significant negative correlation with MRS (*r* = -.268, *CI* = [-.324, -.209]) and ARS (*r* = -.130, *CI* = [-.190, -.069]), and was positively correlated with ERS (*r* = .284, *CI* = [.226, .340]). Thus, the two approaches yielded substantially different trait correlations.

The correlations between self-esteem and RSs under the MGPCM were all significantly different from zero and small (ARS: *r* = .155, *CI* = [.093, .214]; ERS: *r* = .176, *CI* = [.115, .235]; MRS: *r* = -.102, *CI* = [-.163, -.041]). Similarly, the MPT showed significant correlations between self-esteem and RS traits, although they were all negative (MRS: *r* = -.758, *CI* = [-.784, -.731]; ERS: *r* = -.696, *CI* = [-.727, -.663]; ARS: *r* = -.582, *CI* =

[-.621, -.540]). Again, the respective correlations differed substantially between the two approaches and, partly, yielded effects in different directions. Thus, the modeling choice had a substantial impact on the correlation between the content and RS traits. This may be a result of the low reliabilities of the RS traits, especially for the RSES (cf. Table 1) and the fact that the MGPCM had latent variances fixed to 1 for model identification whereas they were estimated freely in the MPT and differed markedly from 1 (*Var* = 6.67 for RSES, and *Var* = 2.74 for NFCS; see Table 1).

## Correlations with Validity Variables

The correlations between NFCS and RSES ranged from *r* = .126 to *r* = .180, which was in the lower range of findings in the literature. All confidence intervals around the pairwise differences included zero (Table 3). The correlations with gender were small (*r* = -.196 and *r* = -.134, NFCS; *r* = -.084 and *r* = -.045, RSES), which was consistent with findings in the literature. The correlation coefficients of the NFCS differed significantly between MPT and SEM, no differences were significant for the RSES (Table 3). Similarly, the correlation coefficients with age did not differ significantly (NFCS: *r* = [-.018; .019]; RSES: *r* = [.108; .154]; Table 3), which both reflected previous findings. While the regression coefficients for education varied between the models for both NFCS and RSES (Table 4), the differences were not significant for both educational levels. The effect sizes were also in line with previous findings.

The correlations with the Big Five were noticeably different between the models, but fell in the general range of previous findings. Table S6 in the online supplement contains the correlation coefficients. Across all facets and for both RSES (except conscientiousness) and NFCS, the correlations under the SEM were significantly different from the correlations under the other models. Additionally, the correlations with emotional stability and openness differed significantly between GPCM, MGPCM and MPT (Table S6).

## Discussion

Psychological research is dominated by self-report measurements. However, item responses in these instruments can be distorted by different RSs reflecting systematic behavioral tendencies beside the target trait (e.g., Baumgartner & Steenkamp, 2001). Although different statistical approaches have been developed to correct for RSs (e.g., Falk & Cai, 2016; Plieninger & Heck, 2018), these operationalize the latent trait differently. Therefore, content traits corrected for RSs might represent different constructs depending on the correction method. The present study evaluated the effect of different correction methods on the external validity of the content traits. We evaluated how RSES and NFCS correlated with each other, the measured RSs and with external criterion variables using three popular methods of correcting for RSs. We found that, following from similar correlational structures throughout the study, the four investigated models estimate similar, but different personality traits. While the content traits seem adequately comparable, the RS traits differ in their relationship to the target construct. The correlations were generally smaller under the MGPCM than the MPT. Additionally, they sometimes even differed in the direction of the correlations. Thus, depending on the chosen correction method, researchers might reach different conclusions regarding the nomological net of RSs. This might stem from several factors. First, the variances were fixed to 1 in the MGPCMs, but varied freely in the MPTs. Second, the MPTs were based on one parameter IRT models and estimated RS propensity parameters for the items, whereas the MGPCM included a discrimination parameter per latent trait and item, but no RS propensity parameter.

The correlations of the content traits with criterion variables reproduced basic effects from the literature. For example, NFCS and RSES were positively correlated with each other and education, but insubstantially with age or gender (Cacioppo et al., 1996; Sinclair et al., 2010). The differences between corrected and even uncorrected scores were small and indicated comparable validities if the external variables were not subject to response styles. Accordingly, the correlations with the Big Five, which were not corrected for RSs, varied substantially, especially between SEM and IRT methods. Although the IRT models seemed more robust towards distorted responses, all variables in an analysis that could be subject to RSs should be corrected to ensure unbiased results.

Moreover, there were differences in the effect sizes although all methods recovered the same correlational patterns if RSs were treated properly. The true effect sizes might, therefore, not be found with one single model, but a model averaging framework such as Bayesian model averaging (Hoeting et al., 1999) or a maximum likelihood version thereof (Lu et al., 2015).

## Limitations

The RSES response data was severely skewed towards higher values of self-esteem, which led us to combine the extreme categories of the scale, effectively creating a 5-point scale. The MPT’s sensitivity to the skewness manifested in the extreme variance estimates. The same can be concluded from the reliability estimates that were quite low for the RSES when treated with the complex MPT and MGPCM, more so than for the NFCS. Thus, only the 18-item NFCS seemed to have contained enough information to distinguish the different person traits, which might be disputed for the RSES.

Moreover, because this study used empirical data, the true effects were unknown and, therefore, the bias in correction methods could not be assessed. We used two personality scales in an attempt to generalize our findings. The investigated models showed similar behavior across both scales. This was similar to other studies examining the effect of RS correction methods. Therefore, we are confident in the robustness of the presented findings.

## Conclusions

This study investigated the impact of different approaches to correct for RSs. The IRT based methods yielded comparable results even for criterion variables subject to RSs, while the effect sizes of the SEM differed significantly in this case. All variables should, thus, be treated for RSs. Furthermore, it may be advisable to first investigate the presence of RSs in personality inventories with a MIRT model befitting the items at hand (e.g., the MPT would be better suited for item formulations that are theoretically prone to RSs than the MGPCM without specific RS elicitation parameters) and then, if RSs are not relevant, use a simpler model like the GPCM. Furthermore, the differences in the effect sizes might warrant an ensemble approach instead of a single correction method.

## References

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181. https://doi.org/10.1111/bmsp.12086

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197-253. https://doi.org/10.1037/0033-2909.119.2.197

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, *10*(4), 1-12. https://doi.org/10.1371/journal.pone.0121945

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328-347. https://doi.org/10.1037/met0000059

Franck, E., De Raedt, R., Barbez, C., & Rosseel, Y. (2008). Psychometric properties of the Dutch Rosenberg self-esteem scale. *Psychologica Belgica, 48*(1), 25-35. https://doi.org/10.5334/pb-48-1-25

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, *7*(1), 7-28.

Greenleaf, E. A. (1992). Measuring extreme response style. *Public Opinion Quarterly, 56*(3), 328–351. https://doi.org/10.1086/269326

He, J., Bartram, D., Inceoglu, I., & van de Vijver, F. J. R. (2014). Response styles and personality traits: A multilevel analysis. *Journal of Cross-Cultural Psychology, 45*(7), 1028–1045. https://doi.org/10.1177/0022022114534773

Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 1): A model integration. *Psychological Methods*, *25(*5), 560–576. https://doi.org/10.1037/met0000249

Hoeting, J., Madigan, D., Raftery, A., & Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*(4), 382-401. https://doi.org/10.1214/ss/1009212519

Kieruj, N. D., & Moors, G. (2011). Response style behavior: question format dependent or personal style? *Quality & Quantity, 47*(1), 193–211. https://www.doi.org/10.1007/s11135-011-9511-4

Kutscher, T., & Eid, M. (2020). The Effect of Rating Scale Length on the Occurrence of Inappropriate Category Use for the Assessment of Job Satisfaction: an Experimental Online Study. *Journal of Well-Being Assessment, 4*(1), 1–35. https://www.doi.org/10.1007/s41543-020-00024-2

Liu, M., Lee, S., & Conrad, F. G. (2015). Comparing extreme response styles between agree-disagree and item-specific scales, *Public Opinion Quarterly*, *79*(4), 952–975. https://doi.org/10.1093/poq/nfv034

Lu, D., Ye, M., & Curtis, G. P. (2015). Maximum likelihood Bayesian model averaging and its predictive analysis for groundwater reactive transport models. *Journal of Hydrology*, *529*, 1859–1873. https://doi.org/10.1016/j.jhydrol.2015.07.029

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical and Statistical Psychology, 72*(3), 501–516. https://doi.org/10.1111/bmsp.12158

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16(2), 159–176*. https://doi.org/10.1177/014662169201600206

Plieninger, H. (2021). Developing and Applying IR-Tree Models: Guidelines, Caveats, and an Extension to Multiple Groups. *Organizational Research Methods, 24*(3), 654–670. https://www.doi.org/10.1177/1094428120911096

Plieninger, H., & Heck, D. W. (2018). A new model for acquiescence at the interface of psychometrics and cognitive psychology. *Multivariate Behavioral Research*, *53*(5), 633–654. https://doi.org/10.1080/00273171.2018.1469966

Primi, R., De Fruyt, F., Santos, D., Antonoplis, S., & John, O. P. (2020). True or false? Keying direction and acquiescence influence the validity of socio-emotional skills items in predicting high school achievement. *International Journal of Testing, 20*(2), 97-121. https://doi.org/10.1080/15305058.2019.1673398

Primi, R., Santos, D., De Fruyt, F., & John, O.P. (2019), Comparison of classical and modern methods for measuring and correcting for acquiescence. *British Journal Mathematical and Statistical Psychology, 72*(3), 447-465. https://doi.org/10.1111/bmsp.12168

Pullmann, H., & Allik, J. (2000). The Rosenberg Self-Esteem Scale: its dimensionality, stability and personality correlates in Estonian. *Personality and Individual Differences, 28*(4), 701–715. https://www.doi.org/10.1016/s0191-8869(99)00132-4

R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer. https://doi.org/10.1007/978-0-387-89976-3

Revelle, W. (2020). psych: Procedures for Personality and Psychological Research [Computer software]. Northwestern University.

Robitzsch, A., Kiefer, T., & Wu, M. (2021). TAM: Test Analysis Modules [Computer software]. https://CRAN.R-project.org/package=TAM

Rosenberg, M. (1965). *Society and the adolescent self-image. Princeton University Press.*

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

Scherpenzeel, A. C., & Das, M. (2010). True longitudinal and probability-based internet panels: Evidence from the Netherlands. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies* (1st ed., pp. 77-104). https://doi.org/10.4324/9780203844922-4

Schmitt, D. P., & Allik, J. (2005). Simultaneous Administration of the Rosenberg Self-Esteem Scale in 53 Nations: Exploring the Universal and Culture-Specific Features of Global Self-Esteem. *Journal of Personality and Social Psychology, 89*(4), 623–642. https://www.doi.org/10.1037/0022-3514.89.4.623

Sinclair, S. J., Blais, M. A., Gansler, D. A., Sandberg, E., Bistis, K., & LoCicero, A. (2010). Psychometric properties of the Rosenberg Self-Esteem Scale: Overall and across demographic groups living within the United States. *Evaluation & the Health Professions, 33*(1), 56–80. https://doi.org/10.1177/0163278709356187

Soto, C. J., & John, O. P. (2019). Optimizing the length, width, and balance of a personality scale: How do internal characteristics affect external validity? *Psychological Assessment, 31*(4), 444–459. https://doi.org/10.1037/pas0000586

Stan Development Team (2021). RStan: the R interface to Stan [Computer software]. https://mc-stan.org

Van Vaerenbergh, Y., & Thomas, T. D. (2013). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. https://doi.org/10.1093/ijpor/eds021

Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing, 27*(3), 236–247. https://www.doi.org/10.1016/j.ijresmar.2010.02.004

Wetzel, E., Böhnke, J. R., & Rose, N. (2016). A simulation study on methods of correcting for the effects of extreme response style. *Educational and Psychological Measurement, 76*(2), 304–324. https://doi.org/10.1177/0013164415591848

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment, 33*, 352-364. https://doi.org/10.1027/1015-5759/a000291

Zhang, Y., & Wang, Y. (2020). Validity of three IRT models for measuring and controlling extreme and midpoint response styles. *Frontiers in Psychology, 11,* 271. https://doi.org/10.3389/fpsyg.2020.00271

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399-413. https://doi.org/10.1037/1082-989x.12.4.399

**Table 1**

*Reliability and Variance Estimates of the Latent Traits*

| | *Reliability* | | | | *Variance* | | | |
|---|---|---|---|---|---|---|---|---|
| *Model* | *Trait* | *MRS* | *ARS* | *ERS* | *Trait* | *MRS* | *ARS* | *ERS* |
| *RSES* | | | | | | | | |
| GPCM | 0.792 | | | | 0.967 | | | |
| SEM | 0.843 | | | | 0.997 | | | |
| MPT | 0.598 | 0.739 | 0.613 | 0.797 | 6.666 | 1.021 | 1.787 | 1.526 |
| MGPCM | 0.669 | 0.303 | 0.347 | 0.686 | 1.013 | 1.046 | 1.004 | 1.050 |
| *NFCS* | | | | | | | | |
| GPCM | 0.905 | | | | 0.989 | | | |
| SEM | 0.865 | | | | 1.001 | | | |
| MPT | 0.787 | 0.770 | 0.558 | 0.569 | 2.736 | 0.444 | 0.595 | 0.105 |
| MGPCM | 0.861 | 0.668 | 0.617 | 0.724 | 1.028 | 0.975 | 0.970 | 1.022 |

*Notes.* NFCS = Need for Cognition scale, RSES = Rosenberg Self-Esteem Scale, (M)GPCM = (multidimensional) generalized partial credit model (variances set to 1 in the models), MPT = multinomial processing tree, SEM = structural equation model with structured residuals, ERS = extreme response style, ARS = acquiescence response style, MRS = midpoint response style

**Table 2**

*Correlation Coefficients of Latent Content and Response Style Traits*

| *MGPCM* | | | | *MPT* | | | |
|---|---|---|---|---|---|---|---|
| Traits | *r* | l-CI | u-CI | Traits | *r* | l-CI | u-CI |
| *RSES* | | | | *RSES* | | | |
| RSES, MRS | -0.102 | -0.163 | -0.041 | RSES, MRS | -0.758 | -0.784 | -0.731 |
| RSES, ERS | 0.176 | 0.115 | 0.235 | RSES, ERS | -0.696 | -0.727 | -0.663 |
| RSES, ARS | 0.155 | 0.093 | 0.214 | RSES, ARS | -0.582 | -0.621 | -0.540 |
| MRS, ERS | -0.071 | -0.133 | -0.009 | MRS, ERS | 0.853 | 0.836 | 0.869 |
| MRS, ARS | -0.009 | -0.071 | 0.053 | MRS, ARS | 0.619 | 0.579 | 0.656 |
| ARS, ERS | 0.116 | 0.054 | 0.176 | ARS, ERS | 0.673 | 0.638 | 0.706 |
| *NFCS* | | | | *NFCS* | | | |
| NFCS, MRS | -0.107 | -0.168 | -0.045 | NFCS, MRS | -0.268 | -0.324 | -0.209 |
| NFCS, ERS | 0.050 | -0.012 | 0.112 | NFCS, ERS | 0.284 | 0.226 | 0.340 |
| NFCS, ARS | 0.080 | 0.018 | 0.141 | NFCS, ARS | -0.130 | -0.190 | -0.069 |
| MRS, ERS | 0.003 | -0.059 | 0.064 | MRS, ERS | -0.302 | -0.357 | -0.245 |
| MRS, ARS | 0.089 | 0.027 | 0.150 | MRS, ARS | 0.073 | 0.011 | 0.134 |
| ARS, ERS | 0.145 | 0.084 | 0.205 | ARS, ERS | 0.393 | 0.340 | 0.444 |

*Note.* NFCS = Need for Cognition scale, RSES = Rosenberg Self-Esteem Scale, MGPCM = multidimensional generalized partial credit model, MPT = multinomial processing tree, *r* = correlation coefficient, l-CI/u-CI = boundaries of confidence interval created at alpha level 0.05, MRS = midpoint response style, ERS = extreme response style, ARS = acquiescence response style

**Table 3**

*Correlation Coefficients of Target Traits and Validity Variables Compared Across Models*

| *Model 1* | *Model 2* | $r_1$ | $r_2$ | *diff* | *N* | *l-CI* | *u-CI* |
|---|---|---|---|---|---|---|---|
| | | | *NFCS and RSES* | | | | |
| MGPCM | GPCM | 0.126 | .180 | -0.054 | 1000 | -0.113 | 0.005 |
| MGPCM | MPT | 0.126 | .131 | -0.005 | 1000 | -0.071 | 0.061 |
| MGPCM | SEM | 0.126 | .173 | -0.046 | 1000 | -0.106 | 0.013 |
| GPCM | MPT | 0.180 | .131 | 0.049 | 1000 | -0.013 | 0.111 |
| GPCM | SEM | 0.180 | .173 | 0.008 | 1000 | -0.046 | 0.062 |
| MPT | SEM | 0.131 | .173 | -0.041 | 1000 | -0.104 | 0.022 |
| | | | *Gender* | | | | |
| *NFCS* | | | | | | | |
| MGPCM | GPCM | -0.168 | -0.176 | 0.008 | 990 | -0.028 | 0.044 |
| MGPCM | MPT | -0.168 | -0.134 | -0.034 | 990 | -0.077 | 0.009 |
| MGPCM | SEM | -0.168 | -0.196 | 0.028 | 990 | -0.013 | 0.068 |
| GPCM | MPT | -0.176 | -0.134 | -0.041 | 990 | -0.085 | 0.003 |
| GPCM | SEM | -0.176 | -0.196 | 0.020 | 990 | -0.015 | 0.055 |
| MPT | SEM | -0.134 | -0.196 | 0.061 | 990 | 0.015 | 0.108 |
| *RSES* | | | | | | | |
| MGPCM | GPCM | -0.081 | -0.063 | -0.018 | 990 | -0.071 | 0.035 |
| MGPCM | MPT | -0.081 | -0.045 | -0.036 | 990 | -0.094 | 0.023 |
| MGPCM | SEM | -0.081 | -0.084 | 0.004 | 990 | -0.049 | 0.056 |
| GPCM | MPT | -0.063 | -0.045 | -0.017 | 990 | -0.070 | 0.035 |
| GPCM | SEM | -0.063 | -0.084 | 0.022 | 990 | -0.028 | 0.071 |
| MPT | SEM | -0.045 | -0.084 | 0.039 | 990 | -0.015 | 0.093 |
| | | | *Age* | | | | |
| *NFCS* | | | | | | | |
| MGPCM | GPCM | -0.012 | 0.008 | -0.020 | 990 | -0.056 | 0.017 |
| MGPCM | MPT | -0.012 | -0.018 | 0.007 | 990 | -0.037 | 0.050 |
| MGPCM | SEM | -0.012 | 0.019 | -0.031 | 990 | -0.071 | 0.010 |
| GPCM | MPT | 0.008 | -0.018 | 0.027 | 990 | -0.018 | 0.071 |
| GPCM | SEM | 0.008 | 0.019 | -0.011 | 990 | -0.046 | 0.025 |
| MPT | SEM | -0.018 | 0.019 | -0.037 | 990 | -0.084 | 0.010 |
| *RSES* | | | | | | | |
| MGPCM | GPCM | 0.117 | 0.108 | 0.009 | 990 | -0.043 | 0.061 |
| MGPCM | MPT | 0.117 | 0.116 | 0.001 | 990 | -0.057 | 0.059 |
| MGPCM | SEM | 0.117 | 0.154 | -0.037 | 990 | -0.089 | 0.016 |
| GPCM | MPT | 0.108 | 0.116 | -0.008 | 990 | -0.060 | 0.044 |
| GPCM | SEM | 0.108 | 0.154 | -0.046 | 990 | -0.095 | 0.004 |
| MPT | SEM | 0.116 | 0.154 | -0.038 | 990 | -0.091 | 0.016 |

*Notes.* NFCS = Need for Cognition scale, RSES = Rosenberg Self-Esteem Scale, (M)GPCM = (multidimensional) generalized partial credit model, MPT = multinomial processing tree, SEM = structural equation model with structured residuals, $r_{1/2}$ = correlation coefficient for the model$_{1/2}$, *diff* = $r_1$-$r_2$, l-CI/u-CI = boundaries of confidence interval around *diff* at alpha level 0.05/6 (Zou, 2007)

**Table 4**

*Regression Coefficients of the Latent Traits and Education*

| *Model 1* | *Model 2* | $b_1$ | $b_2$ | *diff* | $SE_1$ | $SE_2$ | *df* | *z* | *p* |
|---|---|---|---|---|---|---|---|---|---|
| *NFCS* | | | | | | | | | |
| *Level 2* | | | | | | | | | |
| MGPCM | GPCM | 0.149 | 0.171 | -0.021 | 0.112 | 0.109 | 987 | -0.138 | 0.762 |
| MGPCM | MPT | 0.149 | 0.127 | 0.023 | 0.112 | 0.182 | 987 | 0.108 | 0.692 |
| MGPCM | SEM | 0.149 | 0.213 | -0.064 | 0.112 | 0.110 | 987 | -0.406 | 0.683 |
| GPCM | MPT | 0.127 | 0.171 | -0.044 | 0.182 | 0.109 | 987 | -0.210 | 0.732 |
| GPCM | SEM | 0.171 | 0.213 | -0.042 | 0.109 | 0.110 | 987 | -0.273 | 0.763 |
| MPT | SEM | 0.127 | 0.213 | -0.086 | 0.182 | 0.110 | 987 | -0.407 | 0.673 |
| *Level 3* | | | | | | | | | |
| MGPCM | GPCM | 0.526 | 0.557 | -0.030 | 0.108 | 0.105 | 987 | -0.199 | 0.699 |
| MGPCM | MPT | 0.526 | 0.729 | -0.203 | 0.108 | 0.177 | 987 | -0.973 | 0.365 |
| MGPCM | SEM | 0.526 | 0.519 | 0.007 | 0.108 | 0.107 | 987 | 0.047 | 0.803 |
| GPCM | MPT | 0.729 | 0.557 | 0.173 | 0.177 | 0.105 | 987 | 0.831 | 0.417 |
| GPCM | SEM | 0.557 | 0.519 | 0.037 | 0.105 | 0.107 | 987 | 0.247 | 0.696 |
| MPT | SEM | 0.729 | 0.519 | 0.210 | 0.177 | 0.107 | 987 | 1.009 | 0.352 |
| *RSES* | | | | | | | | | |
| *Level 2* | | | | | | | | | |
| MGPCM | GPCM | 0.088 | 0.104 | -0.016 | 0.113 | 0.111 | 987 | -0.102 | 0.718 |
| MGPCM | MPT | 0.088 | 0.120 | -0.032 | 0.113 | 0.290 | 987 | -0.101 | 0.533 |
| MGPCM | SEM | 0.088 | 0.097 | -0.009 | 0.113 | 0.113 | 987 | -0.058 | 0.773 |
| GPCM | MPT | 0.120 | 0.104 | 0.015 | 0.290 | 0.111 | 987 | 0.047 | 0.554 |
| GPCM | SEM | 0.104 | 0.097 | 0.007 | 0.111 | 0.113 | 987 | 0.045 | 0.727 |
| MPT | SEM | 0.120 | 0.097 | 0.023 | 0.290 | 0.113 | 987 | 0.072 | 0.571 |
| *Level 3* | | | | | | | | | |
| MGPCM | GPCM | 0.157 | 0.210 | -0.053 | 0.110 | 0.107 | 987 | -0.345 | 0.693 |
| MGPCM | MPT | 0.157 | 0.414 | -0.257 | 0.110 | 0.281 | 987 | -0.843 | 0.442 |
| MGPCM | SEM | 0.157 | 0.116 | 0.040 | 0.110 | 0.109 | 987 | 0.259 | 0.643 |
| GPCM | MPT | 0.414 | 0.210 | 0.204 | 0.281 | 0.107 | 987 | 0.668 | 0.484 |
| GPCM | SEM | 0.210 | 0.116 | 0.093 | 0.107 | 0.109 | 987 | 0.609 | 0.561 |
| MPT | SEM | 0.414 | 0.116 | 0.297 | 0.281 | 0.109 | 987 | 0.978 | 0.395 |

*Note.* Reference category = “no traditional education”, Level 2 = “finished school education”, Level 3 = “higher education”, NFCS = Need for Cognition scale, RSES = Rosenberg Self-Esteem Scale, (M)GPCM = (multidimensional) generalized partial credit model, MPT = multinomial processing tree, SEM = structural equation model with structured residuals, $b_{1/2}$ = regression coefficient of model$_{1/2}$, *diff* = $b_1$-$b_2$, *SE* = standard error of $b_{1/2}$, *df* = degrees of freedom, *z* = z transformed difference, *p* = p-value

## Open Science

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all measures in the study, and all analyses including all tested models. If we use inferential tests, we report exact p values, effect sizes, and 95% confidence or credible intervals.

Open Data: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported results, including codebook if relevant. The data analyzed in this study and the corresponding information is available at https://www.lissdata.nl/.

Open Materials: We confirm that there is sufficient information for an independent researcher to reproduce all of the reported methodology. The analyses were conducted using R (Version 4.0.5; R Core Team, 2021) and the R packages *rstan* (Stan Development Team, 2021), *psych* (Version 2.0.9; Revelle, 2020), *lavaan* (Version 0.6-8; Rosseel, 2012), *cocor* (Version 1.1-3; Diedenhofen, & Musch, 2015), *mirt* (1.34; Chalmers, 2012) and *TAM* (3.7-16; Robitzsch, Kiefer & Wu, 2021). We provide the computer code used to generate the reported results at https://osf.io/jmkqg/.

Preregistration of Studies and Analysis Plans: This study was not preregistered.

The online supplement is also available at https://osf.io/jmkqg/.