The Version of Record of this manuscript has been published and is available in *Structural Equation Modeling: A Multidisciplinary Journal*, Published online on 25 Sep 2025 http://www.tandfonline.com/10.1080/10705511.2025.2555203

Response Styles and Omitted Variable Bias. Detection and Mitigation

Tomasz Żółtak[a]*, Artur Pokropek[b] and Marek Muszyński[c]

[a]*Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland, ORCID: 0000-0003-1354-4472, email: tomasz.zoltak@ifispan.edu.pl*

[b]*Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland, ORCID: 0000-0002-5899-2917*

[c]*Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland, ORCID: 0000-0002-9306-1237*

Funding: This work was supported by the Polish National Science Centre (NCN) under Grant 2019/33/B/HS6/00937.

Disclosure statement: The authors report there are no competing interests to declare.

Response Styles and Omitted Variable Bias. Detection and Mitigation

Response styles (RS) introduce systematic errors in survey data using rating scales, distorting statistical analyses. This study examines their impact in Structural Equation Modeling (SEM), treating RS as a source of omitted variable bias. It evaluates two RS modeling approaches—Item Response Tree (IRTree) and Multidimensional Nominal Response Models (MNRMs) – through simulations and empirical analysis. Results show that correctly applied models improve parameter precision, while misaligned models produce bias and unreliable confidence intervals. Analyzing OECD's PIAAC dataset reveals cross-country RS variability, affecting correlations between self-reported learning strategies and cognitive skills. Explicitly modeling RS enhances model fit, improving the validity of survey-based conclusions.

Keywords: response styles, IRTree models, multidimensional nominal response model, omitted variable bias

Introduction

In social sciences, latent constructs like opinions, attitudes, and values are commonly measured through questionnaires using rating scales. While convenient for quantitative analysis, these scales risk validity issues due to varied response tendencies. A major concern is whether all respondents interpret response categories consistently and answer with regard for substantial content of the questions.

This problem was conceptualized as response styles (RS), defined as construct-irrelevant response patterns (Paulhus, 1991). RS are invalid responses that follow a specific pattern (Khorramdel & von Davier, 2014) but are unrelated with the construct of interest (Rorer, 1965). Among many response styles (Van Vaerenbergh & Thomas, 2013), three are widely studied: extreme response style (ERS), marked by excessive use of extreme response options; midpoint response style (MRS), characterized by the preference for middle response options; and acquiescence response style (ARS), where respondents agree with items

regardless of their content. Response styles (RS) occur regardless of data collection mode (Liu et al., 2017) or question format (Kieruj & Moors, 2013). Notably, RS are linked to traits of interest in social science, including gender (Weijters et al., 2010), cultural differences (Bolt & Newton, 2011; He et al., 2014), education level (Meisenberg & Williams, 2008), age (De Jong et al., 2008), personality traits (He et al., 2014), and motivation (Gibbons et al., 1999). As a result, RS can undermine measurement validity by introducing systematic errors (Van Vaerenbergh & Thomas, 2013).

The issue of response styles (RS) in statistical inference can be seen as omitted variable bias (Clarke, 2005; Steiner & Kim, 2016; Van der Weele, 2022) when RS correlates with both dependent and independent variables but is not explicitly modeled. Figure 1 illustrates the consequences of omitting RS, which can lead to phantom (spurious) effects (Panel a), suppression effects (Panel b), or overestimation of relationships (Panel c). These issues were recognized as early as Cronbach (1946), who highlighted "extraneous" variables that must be controlled to account for measurement errors linked to response formats. Such variance, known as method variance (Podsakoff et al., 2003), can cause both type I and type II errors if unaddressed. RS, a form of method variance, may either inflate or deflate observed relationships (Podsakoff et al., 2003). Alternatively, RS can be viewed as construct-irrelevant variance, posing a fundamental threat to validity (Messick, 1995).

<<Figure 1 about here>>

Given the strong evidence of RS correlations with key predictors in social analyses, the scenarios in Figure 1 are not just possible but highly plausible. However, researchers using rating scales are not powerless. Many omitted variables disrupt inference because they are unobservable, hard to measure, or overlooked during data collection. RS, though not directly observable, can be inferred from response patterns. Once identified, they can be incorporated into models to control for confounding bias caused by RS.

The primary objective of this study is to expand on previous response styles (RS) research by exploring their lesser-known effects on statistical models. Prior studies have focused on RS impacts in basic Confirmatory Factor Analysis (CFA), Item Response Theory (IRT), and simple comparative analyses like t-tests (Baumgartner & Steenkamp, 2001; Moors, 2004; Zhang et al., 2022). We shift focus to Structural Equation Modeling (SEM), a key tool for analyzing inter-variable relationships. Through simulations, we examine the mechanisms through which biases may occur in SEM's structural components while ignoring RS and assess whether controlling for RS in measurement models mitigates these biases. Additionally, we present an empirical analysis using data from 36 entities in the PIAAC study to detect RS, explore country-specific patterns, and evaluate how these techniques affect important study outcomes.

Secondly, our study explores the problem of RS in a new research context – large-scale social studies, in which data come from interviewer-based face-to-face data collection employing many diverse and short, several items long, measurement instruments. This is in contrast to most RS studies concentrating on self-completion data, with long (> 20 items) scales measuring one topic (e.g. personality, Khorramdel & von Davier, 2014).

This paper compares two emerging approaches for detecting and modeling response styles (RS): (1) IRTrees (Böckenholt, 2012), which represent response processes as binary choice sequences, and (2) Multidimensional Nominal Response Models (MNRMs) with predefined RS patterns (Falk & Cai, 2016). We examine whether these frameworks can identify RS in datasets typical of large-scale social studies and compare their detection rates when RS is present. This analysis is key to improving our understanding of data-generation mechanisms in rating scale studies and enhancing the accuracy of statistical models used to analyze them.

Methods for detecting response styles

This section describes two state-of-the-art methods of detecting and modeling RS, focusing on extreme response style (ERS), and midpoint response style (MRS). For simplification, we assume the typical 5-point Likert scale.

IRTrees

The first method presented in the paper was proposed by Partchev and de Boeck (2012) and Böckenholt (2012). In this approach, a response provided using a rating scale is decomposed into multiple response subprocesses represented by binary pseudo-items. The most widely used decomposition for the 5-point scale is depicted in the upper part of Table 1 (for other possible specifications, see: Böckenholt, 2017; Meiser et al., 2019; Plieninger, 2020).

This approach creates three binary pseudo-items: one for midpoint response style (MRS), another for agreement regardless of intensity (AGR), and a third for extreme response style (ERS), independent of agreement or disagreement. No dependencies are imposed between these pseudo-items. This is ensured by assigning missing values to AGR and ERS when a respondent selects the middle category, aligning with standard item response theory (IRT) modeling, which assumes conditional independence of items in estimation.

The binary pseudo-items could be applied to the regular three-dimensional (non-compensatory) multidimensional IRT model. While Böckenholt (2012, 2017) originally used a probit model, below we describe a logistic variant of this model. For the case of a multidimensional two-parameter logistic model (2PLM) with between-item multidimensionality (each pseudo-item loads on only one latent trait but different pseudo-items can load on different traits), the probability of response 1 to pseudo-item *v* that is indicator of latent trait *j* can be defined as:

$$P(Y_v = 1|\theta_j, \beta_v, \alpha_v) = \frac{exp[\alpha_v(\theta_j - \beta_v)]}{1 + exp[\alpha_v(\theta_j - \beta_v)]} \quad (1)$$

where θ_j is a vector of latent trait's values and α_v is the item discrimination for item v on latent trait j (with the restriction that each pseudo-item loads on only one latent trait). Pseudo-items loaded solely by latent variables representing response styles (like MRS and ERS in the example above) are often assumed to have the same discriminations.

This reflects the assumption that, at least within a single battery of items, response styles may be considered stable, item-independent predispositions of respondents (Böckenholt, 2017; Khorramdel & von Davier, 2014). IRTree models are non-compensatory because they represent a series of binary decisions that do not allow high scores on one dimension to compensate for low scores on another, meaning each subprocess (e.g. extreme responding) is treated independently in the analysis.

IRTree models are easy to interpret and have been widely applied to empirical data. Khorramdel and von Davier (2014) demonstrated that controlling for extreme (ERS) and midpoint response styles (MRS) eliminates counter-theory correlations among Big Five personality traits. Plieninger and Meiser (2014) used IRTree decomposition on a four-point reading self-efficacy scale, showing that when ERS was accounted for, the correlation between self-reported reading skills and literacy test performance nearly doubled, revealing a negative suppression effect. Without modeling ERS, this correlation would have been underestimated.

LaHuis et al. (2019) found that adjusting personality measures for response styles increased the correlation between self-reported personality and job performance. Additionally, IRTrees helped resolve the paradoxical negative correlation between questionnaire-measured attitudes and test-measured cognitive proficiency in PISA, which weakened after accounting for response styles (Khorramdel et al., 2017).

<<Table 1 about here>>

In the IRTree decomposition, three pseudo-traits are introduced: midpoint response style (MRS), extreme response style (ERS), and agreement regardless of intensity (AGR). Although our simulation scenarios focus on the consequences of ignoring MRS and ERS, AGR appears naturally in this specification and is part of the standard representation of IRTree models (Böckenholt, 2012). For consistency with this established notation, we retained AGR in the description of the measurement model, but the bias analyses reported in this paper concern only MRS and ERS. In line with Böckenholt's IRTree framework (2012), we define AGR as a trait-related response, that is, a direction decision reflecting endorsement of an item based on the latent trait. This should be clearly distinguished from acquiescent response style (ARS), which represents a content-independent tendency to agree with items irrespective of their meaning (Plieninger & Heck, 2018).

Multidimensional Nominal Response Models for Response Styles

Another approach to modeling RS is Multidimensional Nominal Response Models (MNRM) with a priori specified response styles pattern (Falk & Cai, 2016). Under the MNRM, the probability of a response in category k of item i is modeled as:

$$P(Y_i = k|W, \Theta, b_{ik}) = \frac{exp(b_{ik}+w_{kAGR}\theta_{AGR}+w_{kMRS}\theta_{MRS}+w_{kERS}\theta_{ERS})}{\sum_{h=1}^{5} exp(b_{ih}+w_{hAGR}\theta_{AGR}+w_{hMRS}\theta_{MRS}+w_{hERS}\theta_{ERS})} \quad (2)$$

where k = 1, 2, 3, 4, 5 are possible ordered response categories (for 5-point Likert scale), b_{ik} is item category intercept parameter and $w_{k\circ}$ are weights defined in the scoring matrix depicted in the bottom part of Table 1. Latent trait variances are estimated model parameters in this specification.

Weights in the scoring matrix are designed to ensure clear interpretation of latent variables and proper model identification. Notably, the specification in Table 1 allows for estimating covariances between latent traits, enabling analysis of relationships between the primary trait and response style (for MNRM identification constraints, see Henninger, 2019; Henninger & Meiser, 2020). The MNRM is a compensatory model, meaning high scores on

one latent trait can offset low scores on another. This allows different response styles – such as agreement, midpoint, or extreme responses – to collectively influence the probability of selecting a particular response category.

The MNRM family of models has been used in international large-scale assessments where they enabled adjusting country scores and revealing strong between-country correlations (Falk & Cai, 2016). Adams et al. (2019) demonstrated that MNRM reduces bias, improves measurement precision, and enhances person fit in self-report data. MNRM was further extended to longitudinal data by Deng et al. (2018), who showed that uncorrected extreme response style (ERS) suppressed the predictive power of self-reported affect on smoking cessation. After accounting for ERS, both the mean effect and its variability became significant predictors of smoking behavior, offering new insights into the cessation process.

Simulation study

Data generation models

In the simulation study, we considered three different models generating item responses:

(1) Unidimensional partial credit model (PCM) with no response styles,

(2) Three-dimensional MNRM with middle and extreme response style traits in the form presented in the previous section,

(3) Three-dimensional IRTree model with middle and extreme response style traits in the form presented in the previous section.

For the no-RS model, we assumed a simple linear relationship of the latent trait with the observed dependent variable characterized by a standardized regression coefficient of size δ. However, for models including RS traits we considered two different scenarios (see also Figure S1 in the supplementary materials):

Scenario 1: All modeled latent traits (trait of interest – ToI, MRS, and ERS) had the same linear effect of standardized size δ on the observed dependent variable. The

correlation between MRS and ERS was fixed to zero, but both were correlated with the trait of interest with a correlation equal to -δ.

Scenario 2: Only MRS and ERS had (linear) effect on the observed dependent variable, both with standardized size of δ. The correlation between MRS and ERS was fixed to zero, but both were correlated with the trait being measured with correlation equal to δ.

In both scenarios, we used the same parameter δ both as a standardized effect size and as the correlation coefficient between the trait of interest (ToI) and response styles (RS). This coupling was a deliberate simplification: it prevents independent manipulation of structural paths and correlations, but provides a straightforward didactic setup in which the biasing role of RS omission becomes evident. Scenario 1 (negative correlations) shows how RS can entirely suppress a true effect, while Scenario 2 (positive correlations) shows how RS can create an artificial effect in the absence of any true ToI–outcome relationship. We stress that these scenarios are not meant to reflect realistic empirical structures but rather to serve as boundary conditions, following the tradition of stylized simulation studies in methodological work (e.g., Plieninger, 2017; Henninger, 2020).

Additional factors being manipulated in the simulation study were:

- Effect size δ: 0.4, 0.3, 0.2 or 0.1,
- Number of items in the scale: 3 or 5 – a typical number for unidimensional scale in comparative social research,

Overall, we have considered 40 different simulation conditions that are summarized in Table S1 in online supplementary materials. For each condition, 1000 replications were completed. The number of observations was assumed to be 1000, which is typical for a single country in

comparative large-scale social research (like the European Social Survey or the International Social Survey Programme).

Measurement model

In the no-RS PCM and in the MNRM data-generating models we made the same assumptions regarding model parameters as in previous simulation studies (Henninger, 2020; Plieninger, 2017). Item location (difficulty) parameters were sampled from a truncated-normal distribution TN(0, 1, -1.5, 1.5) and response category thresholds were sampled from a uniform distribution U(-2.5, 2.5).

In the no-RS PCM model, we assumed the latent trait had a variance of 1, while in the three-dimensional MNRM, the agreement (AGR) and midpoint response style (MRS) traits were set to a variance of 1, and the extreme response style (ERS) trait to 4. This decision was based on preliminary analyses of European Social Survey (ESS) Round 9 data, where median variances of MRS and the primary trait were similar, while ERS variance was nearly four times larger (see Table S2 in the supplementary materials for more details).

Finally, we used the Rasch model as a data-generating model for all IRTree scenarios, setting the discrimination parameters in all nodes to be 1 for all items. Item location (difficulty) parameters were sampled from truncated-normal distributions: TN(-1.2, 0.8, -2.8, 0.4) for MRS nodes, TN(-0.4, 2.9, -6.2, 5.4) for Agreement nodes and TN(-1.6, 1.2, -4, 0.8) for ERS nodes with latent trait variances set to 7.6 for agreement (AGR), 4.5 for ERS and 1.7 for MRS, following results of empirical analysis on 5-category response scales included in the ESS9 data (see Table S2).

Data generation procedure

In each iteration, we first generated latent variable values for 1,000 observations, sampling

from either a standard normal distribution (for the no-RS unidimensional model) or a multivariate normal distribution with a predefined covariance matrix. Next, item parameters were sampled, and item responses were generated using the specified measurement model via the R package rstyles (version 0.7.1; Żółtak, 2023). Latent variables were then standardized by dividing by their model-generating standard deviations and scaled by their corresponding regression coefficients according to the simulation condition. The observed dependent variable was computed by summing these weighted latent variables with error terms drawn from a zero-mean normal distribution, ensuring a standard normal distribution of the final dependent variable. Sampling from multivariate and truncated normal distributions was performed using the R package mnormt (version 2.1.1; Azzalini & Genz, 2022). Models were estimated using Mplus (version 8.5), with R packages rstyles and MplusAutomation (version 1.1.0; Hallquist & Wiley, 2018) used for syntax preparation and automation.

Results

Model convergence

If the data-generating model included response styles – either IRTree or MNRM – all models exhibited a 100% convergence rate. However, if the data-generating model was absent of RS, models assuming the existence of RS had considerable convergence problems. The IRTree model converged, depending on the specific condition, in 55-60% of iterations in conditions with 5 items and in 45-50% of iteration in conditions with 3 items while MNRM only in 41-46% of iterations, irrespective the number of items (for detailed results see Table S1 in online supplementary materials).

RS detection

Comparison of the values of the information criteria almost universally led to the selection of the model that corresponded to the data-generating model. Incorrect choices occurred only in conditions with 3 items – slightly more frequent with increasing effect size, but never more

frequently than in 3% of iterations. AIC and BIC led to the same choices if data-generating models included response styles.

Path parameter bias and coverage

Figure 2 presents the bias and RMSE of the path parameter describing the relationship between the trait of interest (ToI) and the dependent variable across different simulation scenarios (see Table S3 in online supplementary materials for more details). As expected, models specified identically to the data-generating process generally produced unbiased estimates with correct confidence interval coverage. The exception was MNRM in scenarios with only 3 items, in which it exhibited a small bias: positive in scenario 1 and negative in scenario 2. The size of this bias was greater the larger was the effect size, nevertheless it was never larger than ±0.06.

Also as expected, ignoring RS led to significant biases. In scenario 1, the true positive relationship between the trait of interest (ToI) and the dependent variable was underestimated, shrinking toward zero. In scenario 2, ignoring RS introduced a false positive effect. When the effect size was at least 0.2, the 95% confidence interval coverage from a simple PCM dropped below 67% in the best case and declined further as the effect size increased; it was also lower in conditions with 5 items than with 3 items. Applying the IRTree model to data generated with MNRM resulted in only a minimally larger bias than the correctly specified model, but with greater estimate variance. On the other hand, applying MNRM to the data generated using the IRTree model led to the overestimation of the path coefficient value in scenario 1. The bias was larger in conditions with 3 items than with 5 items. In scenario 2, the MNRM bias was very small and negative.

In both scenarios, applying a RS model other than the data-generating one resulted in too narrow 95% confidence intervals. The coverage became narrower as the effect size increased. This relation was steeper in scenario 1 (effect size applied to all path coefficients)

than in scenario 2 (effect size applied only to path coefficients of RS), and for the IRTree model compared to the MNRM. Also, the coverage of IRTree estimates was a little worse in conditions with 5 items than in 3 items, while the coverage of MNRM estimates did not depend on this factor.

If the data were generated using a simple PCM, MNRM provided only slightly overestimated path parameters, but with much less precision than PCM and definitely too wide confidence intervals. In the same conditions, the IRTree model estimates were considerably too large, especially in conditions with only 3 items, and very imprecise, but their 95% confidence intervals were only slightly too narrow.

For path parameters describing the relationships between RS and the dependent variable (see Figures S3, S4, and Table S3 in online supplementary materials), the IRTree model produced approximately unbiased estimates when applied to data generated using the same model in both scenarios. When applied to MNRM-generated data, the IRTree model significantly overestimated ERS path coefficients in both scenarios, while MRS bias remained small and comparable to MNRM estimates. Conversely, MNRM applied to MNRM-generated data slightly overestimated path coefficients for both ERS and MRS in conditions with 3 items and larger effect size.

For IRTree-generated data, MNRM estimates were strongly downward biased for ERS and strongly upward biased for MRS in both scenarios. The size of these biases was somewhat larger in conditions with only 3 items. Both models provided good 95% confidence interval coverage when used with data generated by the same model but produced overly narrow confidence intervals when applied to data from the other model.

<<Figure 2 about here>>

Empirical example

Data

For the empirical analysis, we used the publicly available Programme for the International Assessment of Adult Competencies (PIAAC) 2012 dataset from the OECD website, covering 36 entities (see Table S4 for the list of entities and sample sizes). PIAAC (OECD, 2019) is a unique interviewer-based study that links survey data with cognitive tests. Unlike survey responses, test results are unaffected by response styles (RS), making them a suitable dependent variable for assessing RS-induced distortions in correlations. PIAAC measured skills in three domains: literacy, numeracy, and problem-solving.

We focused on a six-item scale measuring the use of elaborate learning strategies, expected to correlate positively with cognitive skills assessed in PIAAC tests (OECD, 2019, pp. 108–109). This tool uses a five-point rating scale with response options labelled: not at all, very little, to some extent, to a high extent, and to a very high extent.

Analysis

For each country and skill combination, we separately fitted SEM models, specifying the use of elaborate learning strategies (self-reported trait of interest, ToI) and response styles (RS) as predictors of cognitive skills. We then compared the path coefficient representing the relationship between learning strategies and skill levels between two models: a baseline model (simple PCM without RS control) and the best-fitting RS-adjusted model, selected based on BIC values. To ensure consistent parameter estimates and standard errors, we applied the unbiased shortcut method (OECD, 2019, pp. 528–529), estimating each model 90 times. For BIC comparisons, we used results from models estimated with the first plausible value and final full sample weights.

Results

Best fitting models

In all 105 cases analyzed, models incorporating RS specifications provided a better fit to the empirical data than the PCM model with no-RS modelled, according to the BIC (see Table S4 for BIC values). With one exception, the best-fitting model was an MNRM controlling for MRS and ERS. The IRTree model fitted best only in the numeracy domain for Finland, where the MNRM model failed to converge in most replications.

Relationship between learning strategies and skills

The results from all 36 samples are summarized in Figure 3 and Table S5. While explicitly modeling response styles improves model fit, it does not always lead to substantial changes in the estimated relationship between the ToI (self-reported use of elaborate learning strategies) and the dependent variable (cognitive skills). In some countries, such as Kazakhstan, Korea, Sweden, Japan, Belgium, and the Netherlands, the estimates remained virtually unchanged. However, in most cases, the standardized path coefficient difference between models ranged from 0.05 to 0.16, with the RS-adjusted model consistently showing a stronger relationship. The largest change occurred in Finland, where the difference exceeded 0.50, but the RS model estimates had extremely large standard errors, making them unreliable. In contrast, some countries exhibited disjoint 95% confidence intervals for path coefficient estimates between the no-RS and best-fitting RS models – a highly conservative criterion for assessing significant differences. This was observed for all cognitive skills for Poland, Canada, France, and Spain. Notably, no relationship was found between the magnitude of coefficient differences and the strength of the relationship in the baseline (no-RS) model.

<<Figure 3 about here>>

The overall picture presented in Figure 3 is clear: in all cases, even when changes are minimal, adjusting for RS results in a higher estimated relationship between the variables of

interest and also higher R-squared statistics (see Table S5). These findings align with results from simulation studies, which suggested that RS may act as a confounding factor that biases parameter estimates. While the impact of RS is less pronounced than in simulations, explicit modelling of RS appears to improve these estimates underscoring the importance of RS control in statistical analyses employing survey data.

Discussion

This study empirically investigated the presence and impact of response styles (RS) in survey research utilizing rating scale items. By employing the Item Response Tree (IRTree) model and the Multidimensional Nominal Response Model (MNRM) we aimed to detect RS and assess the efficacy of these models in contexts typical of large-scale assessments, characterized by five-point rating scales and a limited number of items per construct.

Our findings showed that both the IRTree and MNRM models exhibit a high detection rate for RS, demonstrating their suitability for identifying RS in survey data. The models showed consistently good performance in simulation scenarios with a 5-items instrument and only slightly worse with an extremely short 3-items instrument. This proves their robustness and applicability in practical research settings, at least with sample sizes of about 1000 participants or larger. These models are also easily estimated in popular statistical packages, which is not always the case (cf. Schoenmakers et al., 2024). This high detection capability is crucial because RS can significantly distort survey responses, leading to biased estimates and incorrect inferences if not adequately accounted for (e.g. Khorramdel et al., 2017; Khorramdel & von Davier, 2014).

Before using rating-scales data in statistical analyses, it is advisable to compare model fit between the baseline IRT model and models accounting for RS to empirically verify whether RS are present. This comparison helps identify the model that best represents RS characteristics in a given dataset. Our simulations show that selecting the wrong model can

severely distort parameter estimates. Based on the results, especially BIC values are reliable criteria for model selection. Since different models may yield different conclusions (cf. Schoenmakers et al., 2024), we recommend testing multiple models and selecting the best-fitting one for statistical analysis.

Empirical evidence from the Programme for the International Assessment of Adult Competencies (PIAAC) study, involving data from 36 entities, reinforces these conclusions. In most countries, applying the IRTree and MNRM models for RS adjustment significantly modified the empirical results, leading to more plausible (and theory-backed) relationships between the constructs measured by self-report (here: use of elaborate learning strategies) and the outcomes of interest (here: cognitive skills).

Our simulation results highlight a clear implication for applied research: parameter estimates in models ignoring response styles (RS) can be substantially biased, either underestimating true effects or producing entirely spurious associations. This means that even well-specified structural models may yield misleading substantive conclusions if RS are not accounted for. For practitioners, the consequence is twofold. First, relationships that appear weak or nonsignificant may in fact be suppressed by unmodeled RS. Second, seemingly robust effects may be artefacts of RS rather than genuine substantive relationships. Both situations risk misinforming theory development and policy recommendations. Therefore, before drawing substantive conclusions from rating-scale data, researchers should explicitly test models with and without RS controls, compare their fit, and carefully evaluate whether RS adjustment changes the interpretation of key parameters. Our empirical example illustrates that, in real data, such adjustments frequently strengthen theoretically expected associations, underscoring the practical importance of addressing RS in applied research.

Limitations

Despite the strengths of this study, several limitations should be acknowledged. First, while

the simulation study covered a range of scenarios, it did not encompass all possible conditions encountered in practice. Future research could extend these simulations to include different scale formats, varying numbers of items, and other RS patterns than ERS and MRS, for example acquiescent response style (ARS; Van Vaerenbergh & Thomas, 2013).

Second, our study concentrated only on the within-country consequences of ignoring RS, leaving aside its implications for between-country comparisons (see He et al., 2014; Ulitzsch et al., 2024).

Third, we examined only two of many psychometric models for RS. While these models suit typical social research scenarios, other RS models may be more appropriate in contexts with more response categories or a larger number of items (cf. Henninger, 2020; Henninger & Meiser, 2019).

Finally, our comparison focused solely on models treating RS as a continuous trait. However, a substantial body of research models RS as a categorical latent trait, often using IRT mixture models (e.g., Khorramdel et al., 2019).

Future Directions

Future research could expand on this study's findings in several ways. One promising direction is developing and testing new models or extending existing ones to better capture complex RS patterns. For instance, models that account for multiple RS types simultaneously (e.g. ERS, MRS, but also ARS, cf. Plieninger & Heck, 2018) could enhance accuracy. A key advancement was proposed by Ulitzsch et al. (2023), who introduced an unstructured RS model with minimal assumptions, allowing flexible RS adjustments at both group and individual levels. This approach is particularly useful when the goal is RS adjustment rather than direct investigation. Ulitzsch et al. (2023) demonstrated that this model effectively accounts for multiple RS types in large-scale assessments, where different countries exhibited distinct RS tendencies. In this study, the unstructured RS model also proved more effective in

correcting correlation coefficients. Empirical evidence suggests that RS adjustment has a greater impact when RS differences between groups are larger (e.g., Schoenmakers et al., 2024; Zhang et al., 2022) or when RS is more strongly related to the trait of interest (this study). This highlights the need for further development of flexible RS models that accommodate diverse RS patterns. Future research should validate unstructured RS models in various settings, comparing them to models with different levels of assumed RS structure.

Further studies should also explore the psychological and cultural drivers of RS, offering insights into the variability of RS patterns both between and within groups (Ulitzsch et al., 2023) and even across time within individuals (Merhof & Meiser, 2023). This could inform the design of surveys that are less prone to RS effects. While post-hoc RS adjustment techniques are advancing, research on RS prevention remains limited (Adams et al., 2019; Henninger et al., 2025).

References

Adams, D. J., Bolt, D. M., Deng, S., Smith, S. S., & Baker, T. B. (2019). Using multidimensional item response theory to evaluate how response styles impact measurement. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 466–485.

Azzalini, A., & Genz, A. (2022). *The R package mnormt: The multivariate normal and t distributions (version 2.1.1)*. http://azzalini.stat.unipd.it/SW/Pkg-mnormt/

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.

Böckenholt, U. (2012). Modeling multiple response processes in judgment and choice. *Psychological Methods*, *17*(4), 665–678. https://doi.org/10.1037/a0028111

Böckenholt, U. (2017). Measuring response styles in likert items. *Psychological Methods*, *22*(1), 69–83. https://doi.org/10.1037/met0000106

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*(5), 814–833.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, *22*(4), 341–352.

Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, *6*(4), 475–494. https://doi.org/10.1177/001316444600600405

De Jong, M. G., Steenkamp, J.-B. E., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(1), 104–115.

Deng, S., E. McCarthy, D., E. Piper, M., B. Baker, T., & Bolt, D. M. (2018). Extreme Response Style and the Measurement of Intra-Individual Variability in Affect. *Multivariate Behavioral Research*, *53*(2), 199–218.

Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, *21*(3), 328–347.

Gibbons, J. L., Zellner, J. A., & Rudek, D. J. (1999). Effects of language and meaningfulness on the use of extreme response style by Spanish-English bilinguals. *Cross-Cultural Research*, *33*(4), 369–381.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

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. (2020). A Novel Partial Credit Extension Using Varying Thresholds to Account for Response Tendencies. *Journal of Educational Measurement*, 1–26.

Henninger, M., & Meiser, T. (2020). Different approaches to modeling response styles in divide-by-total item response theory models (part 2): Applications and novel extensions. *Psychological Methods, 25*(5), 577–595. https://doi.org/10.1037/met0000268

Henninger, M., & Plieninger, H. (2020). Different Styles, Different Times: How Response Times Can Inform Our Knowledge About the Response Process in Rating Scale Measurement. *Assessment 28*(5), 1301-1319. https://doi.org/10.1177/1073191119900003

Henninger, M., Plieninger, H., & Meiser, T. (2025). The Effect of Response Formats on Response Style Strength: An Experimental Comparison. European Journal of *Psychological Assessment, 41*(1), 72–88. https://doi.org/10.1027/1015-5759/a000779

Khorramdel, L., & von Davier, M. (2014). Measuring Response Styles Across the Big Five: A Multiscale Extension of an Approach Using Multinomial Processing Trees. *Multivariate Behavioral Research*, *49*(2), 161–177.

Khorramdel, L., Von Davier, M., Bertling, J. P., Roberts, R. D., Kyllonen, P. C., Khorramdel, L., Von Davier, M., Bertling, J. P., Roberts, R. D., & Kyllonen, P. C. (2017). Recent IRT approaches to test and correct for response styles in PISA background questionnaire data: A feasibility study. *Psychological Test and Assessment Modeling*, *59*(1), 71–92.

Khorramdel, L., von Davier, M., & Pokropek, A. (2019). Combining mixture distribution and multidimensional IRTree models for the measurement of extreme response styles. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 538–559.

Kieruj, N. D., & Moors, G. (2013). Response style behavior: Question format dependent or personal style? *Quality and Quantity*, *47*(1), 193–211.

LaHuis, D. M., Blackmore, C. E., Bryant-Lees, K. B., & Delgado, K. (2019). Applying Item

Response Trees to Personality Data in the Selection Context. *Organizational Research Methods*, *22*(4), 1007–1018. https://doi.org/10.1177/1094428118780310

Liu, M., Conrad, F. G., & Lee, S. (2017). Comparing acquiescent and extreme response styles in face-to-face and web surveys. *Quality & Quantity, 51*(2), 941–958. https://doi.org/10.1007/s11135-016-0320-7

Meisenberg, G., & Williams, A. (2008). Are acquiescent and extreme response styles related to low intelligence and education? *Personality and Individual Differences*, *44*(7), 1539–1550.

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.

Merhof, V., & Meiser, T. (2023). Dynamic Response Strategies: Accounting for Response Process Heterogeneity in IRTree Decision Nodes. *Psychometrika*, *88*(4), 1354–1380.

Messick, S. (1995). *Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50*(9), 741–749.

Moors, G. (2004). Facts and artefacts in the comparison of attitudes among ethnic minorities. A multigroup latent class structure model with adjustment for response style behavior. *European Sociological Review*, *20*(4), 303–320.

OECD. (2019). *Technical Report of the Survey of Adult Skills (PIAAC) (3rdedition)*. https://www.oecd.org/skills/piaac/publications/PIAAC_Technical_Report_2019.pdf

Partchev, I., & De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, *40*(1), 23–32. https://doi.org/10.1016/j.intell.2011.11.002

Paulhus, D. L. (1991). Measurement and control of response bias. In Robinson, Shaver, & Wrightsman (Eds.), *Measures of personality and social psychological attitudes*.

Academic Press.

Plieninger, H. (2017). Mountain or Molehill? A Simulation Study on the Impact of Response Styles. *Educational and Psychological Measurement*, *77*(1), 32–53.

Plieninger, H. (2020). Developing and Applying IR-Tree Models: Guidelines, Caveats, and an Extension to Multiple Groups. *Organizational Research Methods*, 1–17.

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.

Plieninger, H., & Meiser, T. (2014). Validity of Multiprocess IRT Models for Separating Content and Response Styles. *Educational and Psychological Measurement*, *74*(5), 875–899. https://doi.org/10.1177/0013164413514998

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, *88*(5), 879–903.

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, *63*(3), 129.

Schoenmakers, M., Tijmstra, J., Vermunt, J., & Bolsinova, M. (2024). Correcting for Extreme Response Style: Model Choice Matters. *Educational and Psychological Measurement*, *84*(1), 145–170. https://doi.org/10.1177/00131644231155838

Steiner, P. M., & Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of Causal Inference*, *4*(2) , 20160009.

Ulitzsch, E., Henninger, M., & Meiser, T. (2024). Differences in response-scale usage are ubiquitous in cross-country comparisons and a potential driver of elusive relationships. *Scientific Reports*, *14*(1), 10890.

Ulitzsch, E., Lüdtke, O., & Robitzsch, A. (2023). The Role of Response Style Adjustments in Cross-Country Comparisons—A Case Study Using Data from the PISA 2015

Questionnaire. *Educational Measurement: Issues and Practice*, *42*(3), 65–79.

Van der Weele, T. J. (2022). Constructed Measures and Causal Inference: Towards a New Model of Measurement for Psychosocial Constructs. *Epidemiology*, *33*(1), 141–151.

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., Geuens, M., & Schillewaert, N. (2010). The Stability of Individual Response Styles. *Psychological Methods*, *15*(1), 96–110. https://doi.org/10.1037/a0018721

Zhang, Y., Yang, Z., & Wang, Y. (2022). The Impact of Extreme Response Style on the Mean Comparison of Two Independent Samples. *SAGE Open*, *12*(2), 215824402211081.

Żółtak, T. (2023). *tzoltak/rstyles: Version 0.7.1* (Version v0.7.1) [Computer software]. https://doi.org/10.5281/ZENODO.7614923

Table 1. Mapping of response categories to the stylized latent traits in IRTree and MNR models used in the simulation study

Decomposition of a 5-point rating scale into binary pseudo-items (BPIs) in IRTree approach			
Category (*k*)	**MRS**	**AGR**	**ERS**
Strongly disagree (1)	0	0	1
Disagree (2)	0	0	0
Neither agree nor disagree (3)	1	-	-
Agree (4)	0	1	0
Strongly agree (5)	0	1	1
Scoring Matrix for Multidimensional Nominal Response Model for 5-point Likert Scale			
Category (*k*)	**AGR**	**MRS**	**ERS**
Strongly disagree (1)	0	0	1
Disagree (2)	1	0	0
Neither agree nor disagree (3)	2	1	0
Agree (4)	3	0	0
Strongly agree (5)	4	0	1

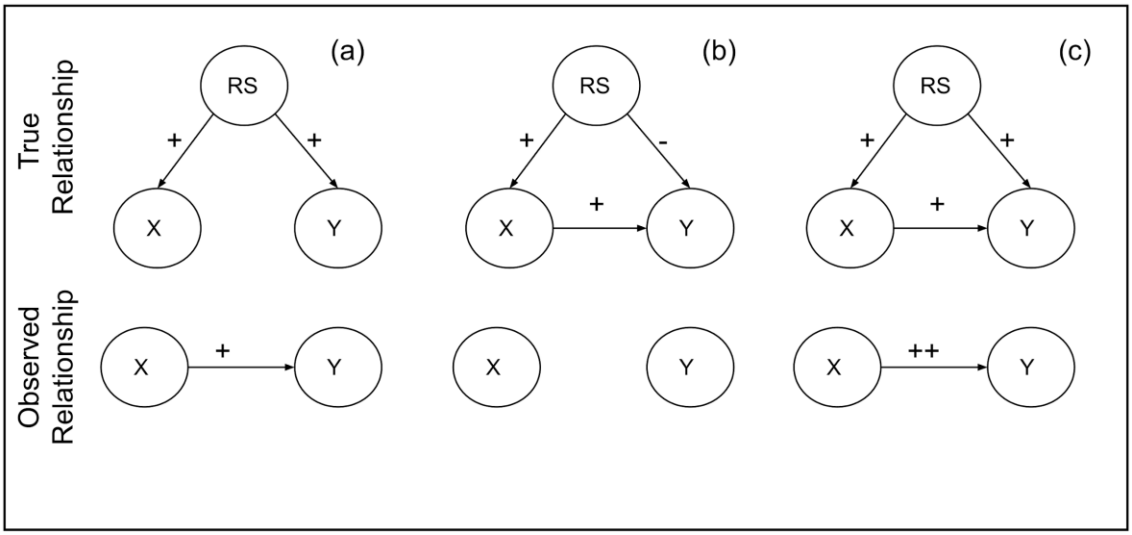

Figure 1. Possible true and observed relationships between dependent (Y) and independent (X) variables while omitting RS.

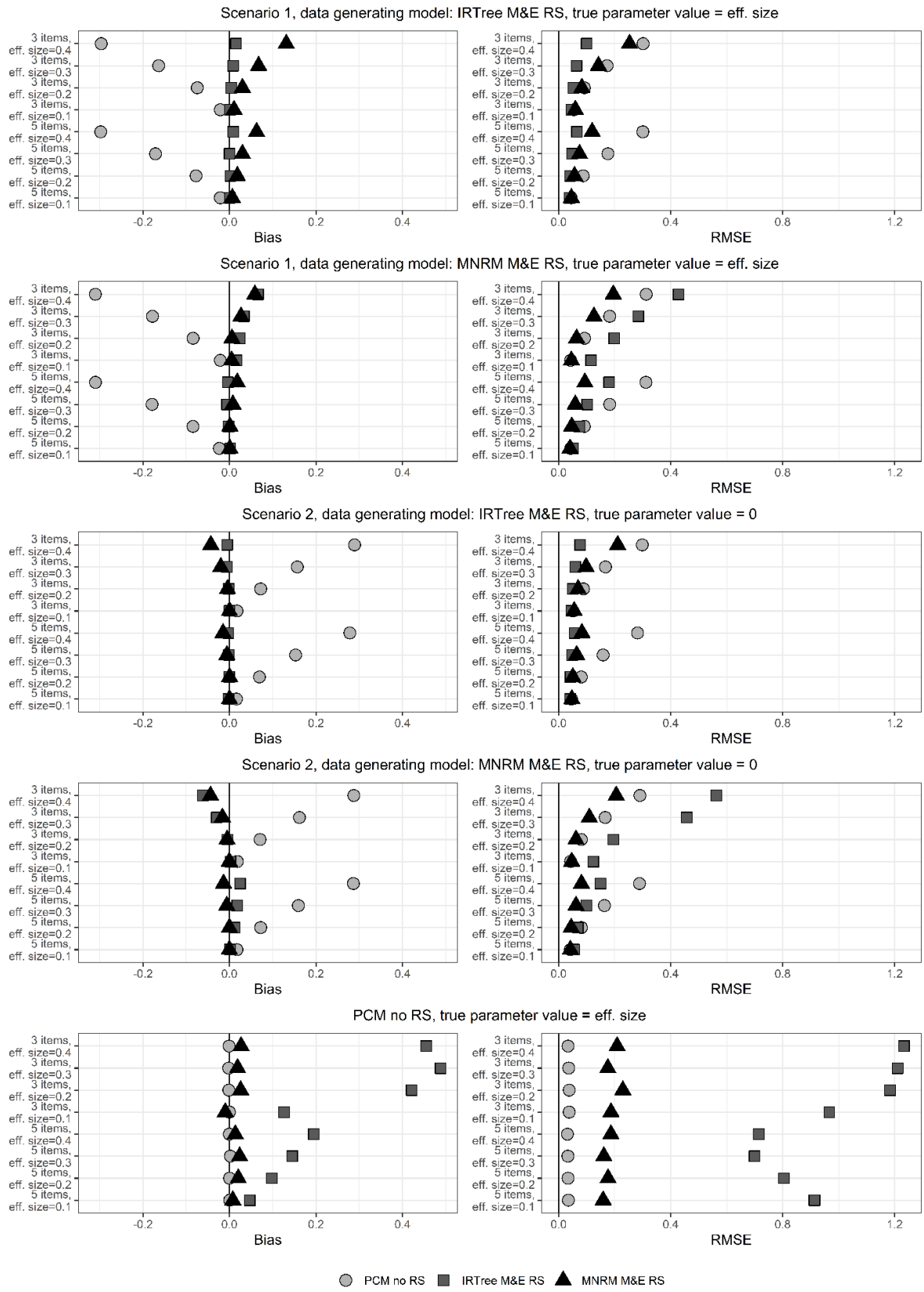

Figure 2. Bias (left panels) and RMSE (right panels) of path coefficients for the trait of interest (ToI) in different simulation scenarios.

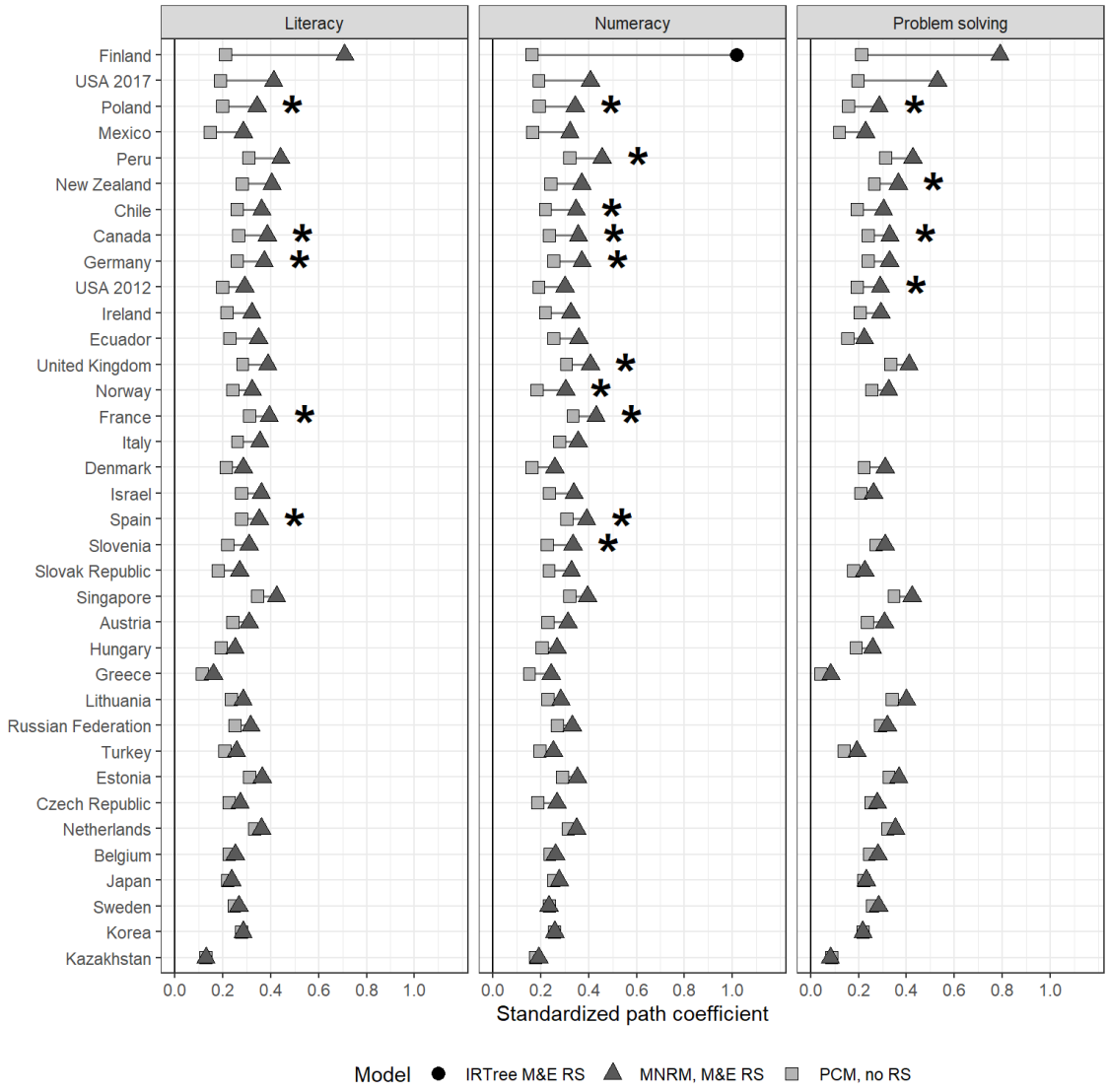

Figure 3. Differences in values of path coefficients describing the relationship between using elaborate learning strategies and skills between best fitting models including response styles and models with no response styles. Estimates with disjoint 95% confidence intervals are marked with stars.

Note. Countries are ordered by the mean difference—across all cognitive skills—between estimates from the best-fitting RS model and the model ignoring RS.