Dissertation

Towards a Deeper Understanding of Response Styles Through Psychometrics

Hansjörg Plieninger

Universität Mannheim

Inaugural dissertation submitted in partial fulfillment of the requirements for the degree Doctor of Social Sciences in the Graduate School of Economic and Social Sciences at the University of Mannheim

November 27, 2017

Supervisor:
Prof. Dr. Thorsten Meiser

Dean of the Faculty of Social Sciences:
Prof. Dr. Michael Diehl

Academic Director of the CDSS:
Prof. Dr. Edgar Erdfelder

Thesis Evaluators:
Prof. Dr. Eunike Wetzel
Prof. Dr. Edgar Erdfelder

Examination Committee:
Prof. Dr. Edgar Erdfelder
Prof. Dr. Eunike Wetzel
Prof. Dr. Thorsten Meiser

Date of Defense:
February 21st, 2018

Für Kathi und Paula und Jakob.

```
> fortunes::fortune("done it.")

It was simple, but you know, it’s always simple
when you’ve done it.
   -- Simone Gabbriellini (after solving a
      problem with a trick suggested on the list)
      R-help (August 2005)
```

Table of Contents

Statement of Originality

Eidesstattliche Versicherung gemäß § 9 Absatz 1 Buchstabe e) der Promotionsordnung der Universität Mannheim zur Erlangung des Doktorgrades der Sozialwissenschaften:

1. Bei der eingereichten Dissertation mit dem Titel „Towards a Deeper Understanding of Response Styles Through Psychometrics“ handelt es sich um mein eigenständig erstelltes eigenes Werk.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtliche Zitate aus anderen Werken als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bisher nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärung bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Abstract

It is well known that respondents answer items not only on the basis of the question content, but also on the basis of their preferences for specific response categories. This phenomenon of so-called response styles has gained a lot of attention in both psychometric and applied work, and research has made steady progress in the last decades. However, there are still many open questions, and selected topics were addressed in three research papers that compose the present, cumulative thesis.

The first paper (Plieninger, 2016) focused on applied settings, where researchers often fear that response styles may threaten the data quality. However, it was unclear how large such biases can be, and this was investigated in simulation studies. Data contaminated by extreme responding and acquiescence were generated from a recently proposed IRT model under a wide range of conditions. Subsequently, the data were analyzed (e.g., Cronbach's alpha, correlations) without controlling response styles, and the resulting bias was investigated. The analyses revealed that bias was small to negligible in many situations, but bias became larger the stronger the correlation between the target trait and response styles was.

The second paper (Plieninger & Heck, 2017) focused on specific psychometric models for response styles, namely, IR-tree models. We showed that these models can be subsumed under the class of hierarchical MPT models. Within this more general framework, we extended an existing model to acquiescence. Simulation studies showed that the Bayesian estimation procedure successfully recovered the parameter values, and an empirical example from personality psychology was used to illustrate the interpretation of the model. Apart from that, comparisons with existing approaches to acquiescence revealed that different concepts of this response style exists, namely, either in terms of a mixture or a shift process, and the proposed model makes it possible to contrast the two accounts.

The third paper (Plieninger, Henninger, & Meiser, 2017) focused on response formats, in particular, the Likert-type format and a recently proposed drag-and-drop format. It was hypothesized that the new format may allow to control response styles as indicated by previous research. We aimed to investigate the underlying mechanisms of this effect as well as possible consequences for reliability and validity. However, a small advantage of the new format over a Likert-type format was only found in one condition where the categories were aligned in two columns. The other conditions, where categories were presented in one column, showed no advantage over the Likert-type format in terms of response styles, reliability, and validity.

In summary, the present thesis has led to a deeper understanding of response styles. Open questions that could not be addressed or were brought up are discussed herein, and routes for future research are described.

Acknowledgements

What follows is a semi-structured list of people that helped me, supported me, shaped my thinking, and taught me interesting and important things. If you appear on this list, I hope that I will have the opportunity to buy you a coffee and warmheartedly thank you in person. If you are not on this list, buy me a coffee and I will tell you how much these people mean to me.

Thorsten Meiser,
Oliver Dickhäuser,
Edgar Erdfelder,
Eunike Wetzel,
Dennis, Dietrich, Florian, Franziska, Gisela, Hanna, Jan, Jana, Maya, Merle, Mirka, Simone,
Isa, Daniel, Ann-Katrin, Pascal, Felix,
Hatice Ecirli, Thomas Gschwend, Hans Jörg Henning, Wolf-Michael Kähler, Bill Revelle, Alexander Robitzsch, Jeffrey Rouder, Hans Christian Waldmann, Otto Walter, Hadley Wickham,
Dominik, Philipp, Judith, Andres, Mechthild, Karin, Traugott, Jakob, Paula, Kathi.

1 Introduction

This cumulative thesis is based on the following three papers:

Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32–53. doi:10.1177/0013164416636655

Plieninger, H. & Heck, D. W. (2017). *A new model for acquiescence at the interface of psychometrics and cognitive psychology.* Manuscript submitted for publication.

Plieninger, H., Henninger, M., & Meiser, T. (2017). *An experimental comparison of the effect of different response formats on response styles.* Manuscript in preparation.

These papers are summarized and discussed in the main part of the present thesis, while details can be found in the respective original work, which is appended. Beforehand, I will give an overview on theoretical and empirical work related to response styles, and I will explain the motivation for the conducted research. In the closing chapter, remaining issues are discussed and ideas for future work emerging from this thesis are outlined.

1.1 Response Styles

Self-reports are a ubiquitous means in social sciences and beyond to elicit ratings of one's personality, attitude, or opinion. The appeal of self-reports lays in the ease of their application and their face validity. However, concerns about taking such ratings at face value have been existing for a long time, and one such concern are response styles. For example, strongly agreeing with an item such as "I am the life of the party" is usually taken as an indication of a high level of extraversion. However, research on response styles has highlighted that a *strongly agree*-response may also be the result of a moderate level of extraversion in combination with a tendency towards extreme responses. Corresponding literature on response styles will be briefly summarized in the following sections, while more comprehensive overviews can be found in the three papers as well as in Wetzel, Böhnke, and Brown (2016) or Van Vaerenbergh and Thomas (2013).

Interindividual differences in response styles were already described decades ago (e.g., Berg & Collier, 1953; Cronbach, 1942). They are broadly defined as preferences for specific response categories, preferences that are not directly related to the item content (Nunnally, 1978; Paulhus, 1991). The three most prominent response styles are the preference for (or avoidance of) extreme categories called extreme response style (ERS), the preference for the midpoint of a scale, called midpoint response style (MRS), and the preference for affirmative categories, called acquiescence response style (ARS). Response styles can best be described by (continuous) latent variables, and there is ample evidence of the stability of individuals' response styles across time and across content domains (e.g., Danner, Aichholzer, & Rammstedt, 2015; Weijters, Geuens, & Schillewaert, 2010a, 2010b; Wetzel, Carstensen, & Böhnke, 2013; Wetzel, Lüdtke, Zettler, & Böhnke, 2016).

1.2 Psychometric Models for Response Styles

Psychometrics is the scientific discipline that is concerned with the foundations of the measurement of psychological variables. Psychometricians develop statistical models and methods to construct, analyze, and interpret tests, questionnaires, and other tools that are used in various areas such as intelligence testing, psychological assessment, or personality research (e.g., Rust, 2009). Psychometrics has always benefited from a vivid exchange between mathematical and statistical developments on the one hand and applied problems especially of intelligence research on the other hand. But also response style research is a small piece in this puzzle: Researchers faced with response styles often sought advice from psychometrics, and developments in response style research such as the concept of systematic measurement error, understanding of multidimensionality, or specific models like item response tree (IR-tree) models subsequently had an impact on psychometric reasoning beyond response styles.

In the beginning, psychometricians such as Cronbach (1942) as well as other scholars attempted to measure response style using simple descriptive statistics such as means or counts across items. For example, Bachman and O'Malley (1984) counted the number of extreme responses and used this as a measure of ERS. When models of item response theory (IRT), factor analysis, and structural equation modeling (SEM) became more popular, researchers started to develop latent variable models for response style. Early approaches often focused on only one response style. For example, mixture-distribution Rasch models consistently favored a 2-class solution over a 1-class solution with a smaller class of respondents showing ERS and a larger class not (e.g., Meiser & Machunsky, 2008; Rost, Carstensen, & von Davier, 1997; Wetzel et al., 2013). ARS, in contrast, was typically found in factor-analysis models that, in addition to some target trait(s), accounted for shared variance among regular and reverse-coded items (e.g., Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Mirowsky & Ross, 1991). Psychometric models published in recent years often incorporate multiple response styles at once, for example, both ERS and ARS. They can be distinguished on different dimensions, for example, regarding the underlying model (e.g., partial credit or nominal response model), whether they model response styles in an exploratory or confirmatory way, whether they allow for content–style correlations or not, and whether they focus on extensions of the person or the threshold parameters (e.g., Bolt & Johnson, 2009; Falk & Cai, 2016; Jin & Wang, 2014; Johnson & Bolt, 2010; Wang, Wilson, & Shih, 2006; Wetzel & Carstensen, 2017).

Research on response style focuses on three different goals: First, on a *psychometric level*, statistical models are developed to make it possible to measure response styles in the first place, at best in a parsimonious and theoretically meaningful way. Second, on a *substantive level*, one is interested in describing and explaining response styles, for example, from an individual-differences or cognitive perspective. Third, on the *applied level*, researchers are not interested in response styles per se, but in purifying a target trait from potentially detrimental response style influences. These three levels mutually reinforce each other, and many models can be used to accomplish several goals simultaneously.

1.3 Current Understanding of Response Styles

Despite the advances that have been made in the past decades, research in recent years has also highlighted open questions on the psychometric, substantive, and applied level (e.g., Van Vaerenbergh & Thomas, 2013; Wetzel, Böhnke, & Brown, 2016), and I will describe some recurring themes in the following. First, comparisons of response style models often focus only on closely related models, but more comprehensive comparisons would help to identify the relative similarities, merits, and weaknesses of each approach. For example, many models for ARS have been proposed (e.g., Billiet & McClendon, 2000; Falk & Cai, 2016; Ferrando, Morales-Vives, & Lorenzo-Seva, 2016; Johnson & Bolt, 2010; Kam & Zhou, 2015; Maydeu-Olivares & Coffman, 2006; ten Berge, 1999; Wetzel & Carstensen, 2017), but they are rarely compared to each other. Simulations by Savalei and Falk (2014) are a notable exception, and more research is needed to contrast models in terms of their statistical properties, their substantive implications, and their usefulness in applied settings.

Second, even though a lot of studies have investigated response style correlates such as age, sex, education, or personality, the respective evidence is mixed. Moreover, it is unclear whether diverging findings are due to natural fluctuations, sample characteristics, or the employed method. Thus, even though response styles can be measured as stable, trait-like constructs, research as not yet been able to develop a coherent nomological net around its core variables.

Third, choosing a method to measure (and control) response styles is difficult for applied researchers because comprehensive guidelines are missing and traditions vary. For example, latent variable models are very popular in the psychometric literature, and these models correct the latent target trait by means of additional latent response style variables. In cross-cultural psychology, in contrast, a very popular method is ipsatization, that is, correcting the observed responses by means of subtracting each respondent's mean (e.g., Fischer, 2004). But guidance on when which method should be chosen is sparse (but see Savalei & Falk, 2014; Wetzel, Böhnke, & Rose, 2016).

In summary, past research has made steady progress in terms of the measurement of response styles, and comprehensive and flexible models are available today. Moreover, the field has implicitly reached consensus that response styles such as ERS, ARS, and MRS exist, and they it may be beneficial to control them in applied work. Furthermore, there is high interest in correlates of response styles, especially in cross-cultural studies. Nevertheless, routes for future research remain and some specific questions are addressed in the present thesis.

Up to here, this overview described the status quo of response styles from the perspective inside the field. In contrast, a look from outside, namely, from a bibliometric viewpoint may offer additional insights, and this is the perspective taken in the following section.

1.4 About Response Style Research

Herein, I will briefly report on two findings from a bibliometric analysis of response style papers that I conducted. Included in the analyses were 826 articles, namely, all peer-reviewed journal articles with the keyword *response style* that were published in 2016 or earlier according to *Web of Science*.

The first analysis concerned published articles. More and more papers on response styles have been published in recent decades as illustrated in Figure 1.1. When fitting an exponential model to the data from 1980 to 2016, a considerably high growth rate of 9.3 % was revealed with a doubling time of 7.4 years ($R^2 = .90$). This indicates that response styles have gained increasing importance in recent years, and these data suggest that this trend will probably continue. Furthermore, it is interesting to note that out of the 826 journal articles, only 45 were published in the category "Psychology, Mathematical" (of *Journal Citation Reports*, e.g., *Psychometrika*, *Multivariate Behavioral Research*). This highlights that response styles are regarded as relevant in many different, often applied fields. For example, the *Journal of Cross-Cultural Psychology* published most papers, namely, 29. Nevertheless, response styles are also gaining interest on the psychometric level with 16 papers published in corresponding journals in the last three years (see Figure 1.1).

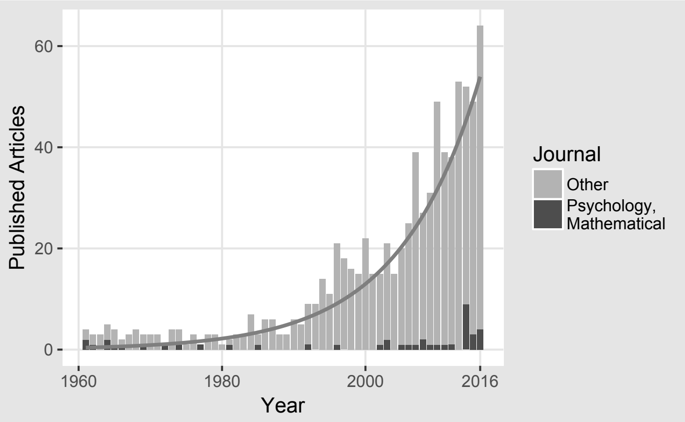


FIGURE 1.1: Bar chart of published journal articles on response styles per year and exponential growth curve.

The second analysis concerned the authors of published papers. It was revealed that, out of all researchers that (co-)authored response style papers, 90 % published only a single response style paper as illustrated in Figure 1.2. Lotka's law of scientific productivity states that the number of authors publishing x papers is related to the number of authors publishing one paper via a specific function, namely, an approximate inverse-square law (Lotka, 1926). For response style articles, this function is significantly steeper ($p = .037$) compared to the usually observed inverse-square law, which is illustrated in Figure 1.2. Thus, the field of response style research has—compared to other areas—more authors that publish only a single paper and fewer authors that publish several papers.

These findings can be interpreted as follows. On the applied level, there is a strong and growing interest in response styles that indicates that researchers feel

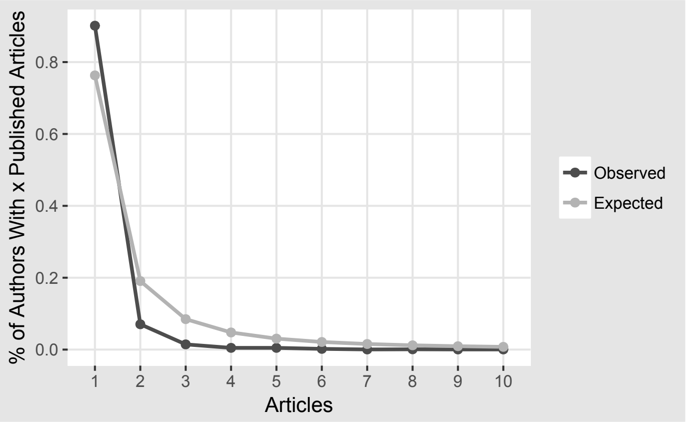


FIGURE 1.2: The number of authors publishing x response style articles deviates from what would be expected under Lotka's law.

the need to take response styles into account. Likewise, there is an increasing interest on the psychometric level. However, there are important, open questions such as comparability of models, substantive meaning of response styles, or best practices, as discussed above. In order to resolve such issues and gain a deeper understanding of response styles, sustained effort is required in terms of dedicated research programs or dissertations. As a result, this will eventually lead to multiple, successive articles published by respective authors—something that is relatively rare in the field of response styles as illustrated in Figure 1.2.

1.5 The Present Research

The aim of the present thesis was to gain a deeper understanding of response styles, and the conducted research focused on three specific topics. The first paper (Plieninger, 2016) addressed the question whether it is at all necessary to control response styles (especially in applied settings). It is often claimed that response styles, if not taken into account, threaten the data quality, but it was unclear how large such biases really are. Therefore, I conducted simulation studies tailored to applied outcomes such as correlations, and I investigated whether ERS and ARS would bias such measures and to what extend. While this paper was mainly targeted at the applied level, it allowed to shed light on the conditions under which response styles are most influential, and this is important for our understanding of response styles on the substantive level. Furthermore, this paper illustrated, on the psychometric level, the use and usefulness of a newly developed IRT model for response styles (Wetzel & Carstensen, 2017).

The second paper (Plieninger & Heck, 2017) was concerned with IR-tree models. These models have recently been proposed by Böckenholt (2012) and De Boeck and Partchev (2012) and quickly became established in the psychometric literature. However, IR-tree models were limited to ERS and MRS. We showed that IR-tree models are special cases of the more general class of hierarchical multinomial processing tree (MPT) models (e.g., Matzke, Dolan, Batchelder, & Wagenmakers, 2015). Within this general framework, we extended an existing IR-tree model to ARS and contrasted it with alternative ARS models. While this paper was mainly targeted at the psychometric level, the developed model will help future research to gain a deeper understanding of ARS on the substantive level.

Up to here, I have limited the discussion of response style control to post-hoc control by means of a statistical method. However, a-priori control, for example, by means of the response format or questionnaire design may be an equally promising approach. In the third paper (Plieninger et al., 2017), we report on the results from two experiments that contrasted the traditional Likert-type response format with a newly developed, so-called drag-and-drop format (Böckenholt, 2017; Thurstone, 1928). We investigated whether the new format may lead to an advantage in terms of response style control, reliability, and validity, and we aimed to delineate the processes that may lead to such an advantage. On the one hand, the new response format may be an alternative means of response style control; and on the other hand, this research may lead to a better understanding of conditions that influence the response process.

2 Towards a Deeper Understanding of Response Style Effects

Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32–53. doi:10.1177/0013164416636655

Part of the reasons to invest resources into research on response styles has always been the claim that response styles may invalidate findings based on questionnaire data. While there was an intensive debate about the importance or negligibility of response styles in the past century (e.g., Bentler, Jackson, & Messick, 1971; Ray, 1979; Rorer, 1965; Schimmack, Böckenholt, & Reisenzein, 2002), this debate subsided in recent years. Nowadays, the claim that response styles have detrimental effects seems to be the mainstream opinion (e.g., Van Vaerenbergh & Thomas, 2013). However, the amount of bias has not been studied systematically and in great detail. Therefore, I conducted a simulation study to (a) investigate the magnitude of bias response styles may induce and (b) identify the conditions under which response styles are most influential. Furthermore, I focused on outcomes relevant for applied researchers, because the claim of bias has of course most impact for applied findings.

2.1 Method

Several simulation studies were designed and carried out using the following procedure: Data contaminated by response styles were generated from a specific IRT model. These data sets were generated under a wide range of conditions, for example, the amount of response style variance was varied. Subsequently, each data set was used, without taking response styles into account, to calculate Cronbach's alpha, scale-score correlation, and individuals' scores. For these three measures, I finally investigated the amount of bias caused by response styles. Details for each of these steps will be given in the following.

A data-generating model had to be chosen that was comprehensive and flexible in order to investigate multiple response styles and a range of conditions. The model proposed by Wetzel and Carstensen (2017) was well suited for the present needs. It is basically a multidimensional partial credit model (MPCM) as, for example, in Adams, Wilson, and Wang (1997). Consider an item with five response categories: Then, in both variants (Adams et al., 1997; Wetzel & Carstensen, 2017), the first person parameter (e.g., extraversion) is multiplied with ordinal weights of $(0, 1, 2, 3, 4)$. In the original MPCM, an item may be indicative also of a second latent variable (e.g., openness), and the same weights are used again. However, Wetzel and Carstensen (2017) proposed to use different weights in order to measure response styles, for example, weights of $(1, 0, 0, 0, 1)$ for ERS and $(0, 0, 0, 1, 1)$ for ARS. Thus, these special weights transform a standard multidimensional model into a response style model that is conceptually similar to existing approaches (e.g., Jin & Wang, 2014; Johnson & Bolt, 2010).

The simulations focused on the effect of response styles on three outcomes, namely, Cronbach's alpha, scale-score correlation (e.g., manifest correlation of extraversion and happiness), and individual scale scores (sum scores). These measures were chosen for two reasons: First, they are heavily used and highly relevant in many fields, and thus it is important to know whether they can be affected by response styles. They can even be conceived as subsequent steps of a research process: Initially, the reliability of a scale is assessed via Cronbach's alpha; subsequently, the validity is studied using correlations; and finally, the scale is used to assign a score to each individual. Second, many other outcomes of potential interest are based on similar concepts, such that these three measures may serve as indicators for other outcomes. For example, confirmatory factor analysis (and SEM) focuses on relationships between items as well as constructs, and this is not inherently different from alpha and scale-score correlation (as measures of relationships among items and constructs, respectively). In other words, if ARS biases manifest scale-score correlations, latent relationships in a SEM will probably be affected in a similar way.

In the simulations, two prominent and qualitatively different response styles were investigated, namely, ARS and ERS. Furthermore, the following independent variables were manipulated: Number of reverse-coded items, response style variance σ^2_{RS}, and correlation ρ between response style and target trait. Pilot simulations revealed that there was virtually no effect of the mean of the response style distribution, of sample size, and of the number of items. Apart from that, five response categories were used, and reasons for not manipulating the number of categories are explained in the appendix of the published paper.

Finally, I would like to highlight some general aspects related to simulation studies. The present simulations could have used three levels for each independent variable, namely, reverse-coded items (e.g., 0, 2, 4), response style variance σ^2_{RS} (e.g., 0, .33, .67), and content–style correlation (e.g., .00, .10, .20) resulting in a $3 \times 3 \times 3$ design. In my opinion (see also Harwell, Stone, Hsu, & Kirisci, 1996), such a procedure would suffer from two issues. First, the independent variable(s) are treated as fixed factors, when one is rather interested in random factors. In other words, there is no difference between (a) running 1,000 replications for each of three factor levels and (b) running 3,000 replications with 3,000 appropriate random values (e.g., σ^2_{RS} sampled from $U(0, 1)$). However, with the latter procedure, it is much easier to detect or rule out interactions and nonlinear effects. The second issue is that the results of such designs are often summarized using only descriptive statistics presented in full-page tables or complex plots. However, they should be treated just like any experiment using appropriate models (like regression) in order to (a) reach a parsimonious and general description of the data, (b) calculate effect sizes, (c) conduct power analyses, (d) detect and describe nonlinear effects, (e) be able to extrapolate, with all due caution, beyond studied conditions, and (f) facilitate interpretation. To address these two issues in the present simulations, the values of the independent variables were randomly drawn from appropriate distributions in each replication. Furthermore, the results were summarized using illustrative plots on the one hand and regression models including interactions and quadratic effects on the other hand.

2.2 Results

The results can be briefly summarized as follows. First, the bias caused by response styles is small or even negligible in many situations. Second, the exception are situations were the target trait and response styles are correlated, and this is worst with respect to ARS in combination with few reverse-coded items. An illustrative overview of the results is presented in Figure 2.1.

More detailed descriptions of the results were obtained by regressing the amount of bias on the manipulated factors and interpretation of the standardized (raw) regression coefficients b^* (b). For example, the effect of ARS on Cronbach's alpha is illustrated in Table 2.1. The intercept of 0.010 indicates that Cronbach's alpha was overestimated by a value of 0.01 in an average condition (e.g., $\rho_{12} = .00$), a negligible amount of bias. But bias increased considerably when the (absolute value) of the content–style correlation ρ_{12} increased and when response style variance increased. Reverse-coded items were a buffer against these effects and reduced bias.

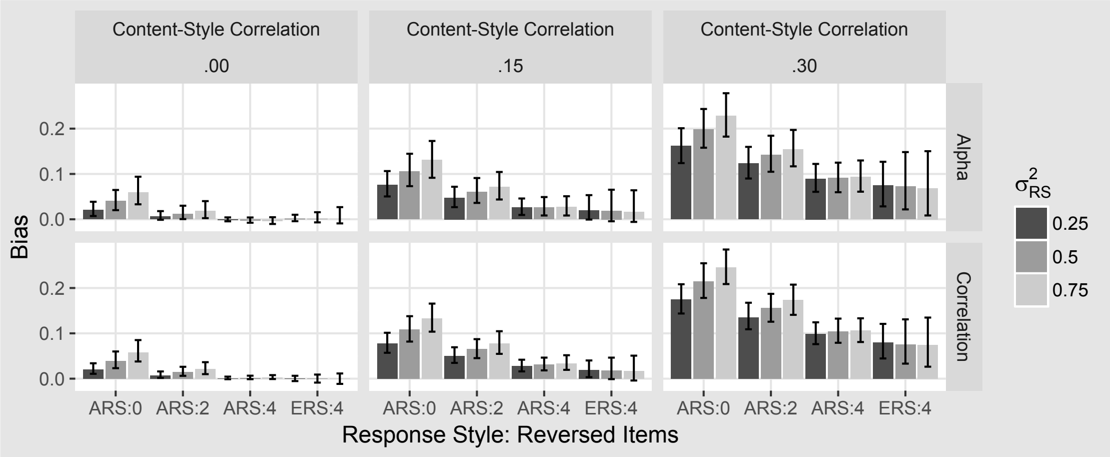


FIGURE 2.1: Overview of bias with respect to Cronbach's alpha (upper panel) and scale-score correlation (lower panel). Displayed is the mean bias as well as percentiles 2.5 and 97.5 across 1,000 replications in each of the selected conditions.

Similar effects were found for the other two outcomes scale-score correlation and individuals' scale scores. ERS caused less bias, and the respective results were comparable to ARS in combination with five reverse-coded items (i.e., a balanced scale). As expected, there was no effect of reverse-coded items for ERS.

TABLE 2.1: Effect of Acquiescence on Bias of Cronbach's Alpha

	b	b^*
(Intercept)	0.010	−0.65
Reversed Items	−0.006	−0.10
σ^2_{ARS}	0.010	0.03
ρ_{12}	0.140	0.38
${\rho_{12}}^2$	0.792	0.65
Reversed $\times\rho_{12}$	−0.055	−0.25
R^2		0.95

Note. All predictor variable were centered. *SE*s of all coefficients were < 0.001.

Finally, the effects caused by response styles will be described in detail for the example of ERS, because the reported results contradict the common impression that response styles lead to severe bias. Oftentimes when dealing with ERS, people think of a person with a moderately high trait level that would "normally" score, for example, $(3, 3, 4, 4, 4, 4, 4, 4, 5, 5)$ but due to high ERS indeed scores $(3, 4, 4, 4, 5, 5, 5, 5, 5, 5)$ on a 10-item scale with five categories. The problem with such examples is threefold: First, such a high level of ERS may occur, but it is quite extreme within prototypical conditions. Second, this effect is only predicted for a specific minority of respondents with the combination of a moderately high trait level a very high ERS level. Third, the same ERS level in combination with a moderately low trait level leads to a decrease in scores: Thus ERS effects on the sample level cancel each other out; ERS induces some error variance but no systematic variance that may lead to systematically biased correlations or the like. These effects are further illustrated in Figure 2.2. The predicted bias of the scale score (sum score) induced by ERS was calculated based on a typical condition of the simulation study reported above (i.e., $\sigma^2_{\text{ERS}} = 0.5$, $\rho_{12} = 0$, 10 items, five categories). The figure illustrates that the bias predicted for the vast majority of persons is (close to) zero and hardly exceeds values of ±2. In other words, ERS may shift a scale score of 40 upwards to 41 or 42 or downwards to 39 or 38, but larger shifts are only predicted for very extreme combination of ERS and target trait. However, as revealed by the results reported above, the predicted shift can be larger if ERS is substantially correlated with the target trait, for extreme values of ERS variance, or if the target trait distribution is not centered around zero (i.e., items are on average too easy or difficult).

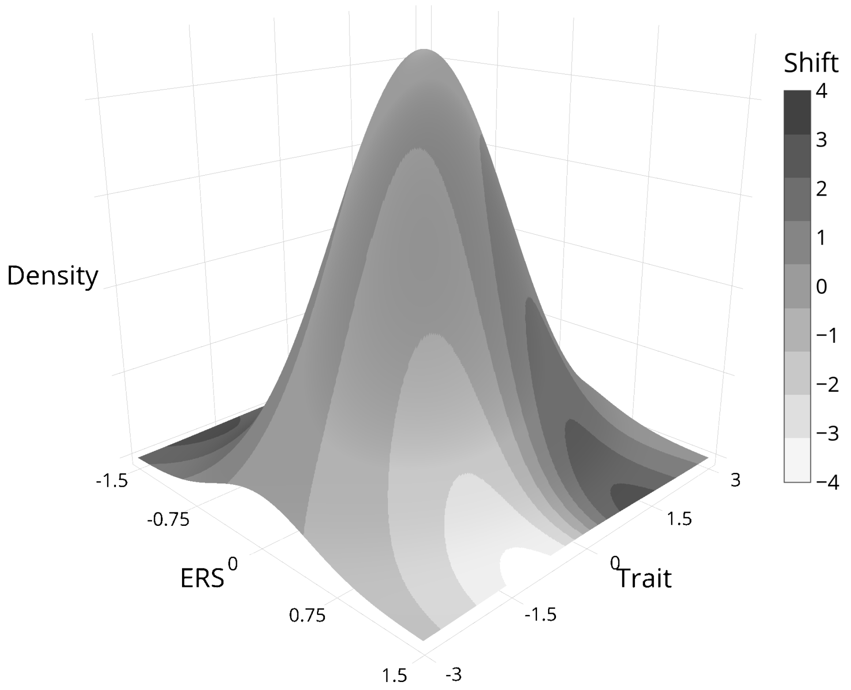


FIGURE 2.2: Density plot of the multivariate normal distribution for target trait ($\sigma^2 = 1$) and ERS ($\sigma^2 = 0.5$). The colors represent the predicted shift of the scale score (possible range from 10 to 50) induced by ERS relative to the absence of ERS. An interactive version may be found at https://plot.ly/~hplieninger/3/.

2.3 Discussion

These results indicate that the detrimental effects of response styles are probably a molehill rather than a mountain in many situations. Serious bias is only expected for substantial content–style correlations, and many respective empirical correlations are small. For example, most correlations with basic personality traits found in the illustrative empirical study in the first paper (Big Five; not reported herein) and the second paper (HEXACO), were $|r| < .10$. However, larger correlations with specific target traits may in principle occur, and individual empirical findings showing that response styles lead to considerable bias may be such cases.

These results help, on the applied level, to identify situations when one should or should not worry about the detrimental effects of response styles. And they allow to quantify the positive effect of reverse-coded items, which have been advocated for a long time (e.g., Cloud & Vaughan, 1970). Moreover, the results highlight, on the substantive level, that it is important to identify correlates of response styles.

While my paper focused on outcomes relevant for applied purposes, other simulation studies carried out at the same time but independently point in the same direction and help complete the picture. Savalei and Falk (2014) investigated the recovery of factor loadings in the presence of ARS and stated that "the 'do nothing' approach [i.e., ignoring ARS] can be surprisingly robust when the ACQ [ARS] factor is not very strong" (p. 407). Furthermore, Wetzel, Böhnke, and Rose (2016) compared different methods to control ERS and stated: "The results of our simulation study imply that ignoring ERS on average hardly affects trait estimates if ERS and the latent trait are uncorrelated or only weakly correlated as typically found in empirical applications" (p. 320). Similar results are found in the papers of Johnson and Bolt (2010) and Ferrando and Lorenzo-Seva (2010) that contain small, illustrative simulation studies.

Thus, it is time to dispel the broad claim and fear that response styles—always and to a large extent—distort questionnaire-based findings. The field should rather move on and focus on other important questions. What is needed, for example, is a better understanding of response styles on the substantive level in order to identify situations in which content and style are likely to be strongly correlated. Furthermore, it is important to know whether the statistical models that are used to control response styles work well in such situations. This would help to provide tailored and accurate guidance to applied researchers who fear that response styles may play a role in their data.

3 Towards a Deeper Understanding of Acquiescence

Plieninger, H. & Heck, D. W. (2017). *A new model for acquiescence at the interface of psychometrics and cognitive psychology.* Manuscript submitted for publication.

Recently, Böckenholt (2012) as well as De Boeck and Partchev (2012) proposed the class of so-called IR-tree models. These models quickly gained interest, each of the two papers has already around 50 citations according to Google Scholar. Within an IR-tree model, a psychologically meaningful tree-like structure of latent processes is assumed to underly the categorical data in question. The models are well suited for response styles, because instead of assuming only one, ordinal response process, they allow to incorporate multiple, distinct processes including response styles.

A response style model for items with five, symmetric response categories (Böckenholt, 2012; De Boeck & Partchev, 2012) is depicted in Figure 3.1 in gray, henceforth called the *Böckenholt Model.* Therein, it is assumed that the response process for respondent i on item j can be described using three stages: An MRS stage is entered with probability m (leading to a midpoint response); a high level of the target trait is reached with probability t (leading to agreement); and an ERS stage is entered with probability e (leading to extreme responses). As can be seen in Figure 3.1, the counter parts of these three stages are entered with the respective counter probabilities. Thereby, the model allows to disentangle three different processes, namely, the target trait and the two response styles ERS and MRS. Another advantage of the model is that it can be fit using standard software for multidimensional IRT models, if each item is recoded into three binary pseudoitems[1]. The model was successfully validated by Plieninger and Meiser (2014), extended within the IR-tree framework (Böckenholt & Meiser, 2017; Khorramdel & von Davier, 2014; Meiser, Plieninger, & Henninger, 2017), and demonstrated

[1] The pseudoitems take on a value of 1 if the outcome of a process was positive, 0 if negative, and missing if not applicable. For example, a response in category 5 is recoded into $(0, 1, 1)$, and a 3 is recoded into $(1, –, –)$.

to be useful in applications (e.g., Zettler, Lang, Hülsheger, & Hilbig, 2016). Very similar approaches were developed by Jeon and De Boeck (2016) or Thissen-Roe and Thissen (2013).

However, estimating IR-tree models based on pseudoitems in general involves "the restriction that each observed response category has a unique path to one of the latent response processes" (Böckenholt, 2012, p. 667). Thus, models are excluded where two paths lead to the same category, for instance, a path t as well as an ARS path that lead to agreement. In other words, the Böckenholt Model cannot accommodate ARS—in contrast to other comprehensive models (e.g., Johnson & Bolt, 2010; Wetzel & Carstensen, 2017). Therefore, our aim was first to demonstrate that IR-tree models are special cases of the more general framework of hierarchical MPT models, and second to develop a model for ARS within this general framework.

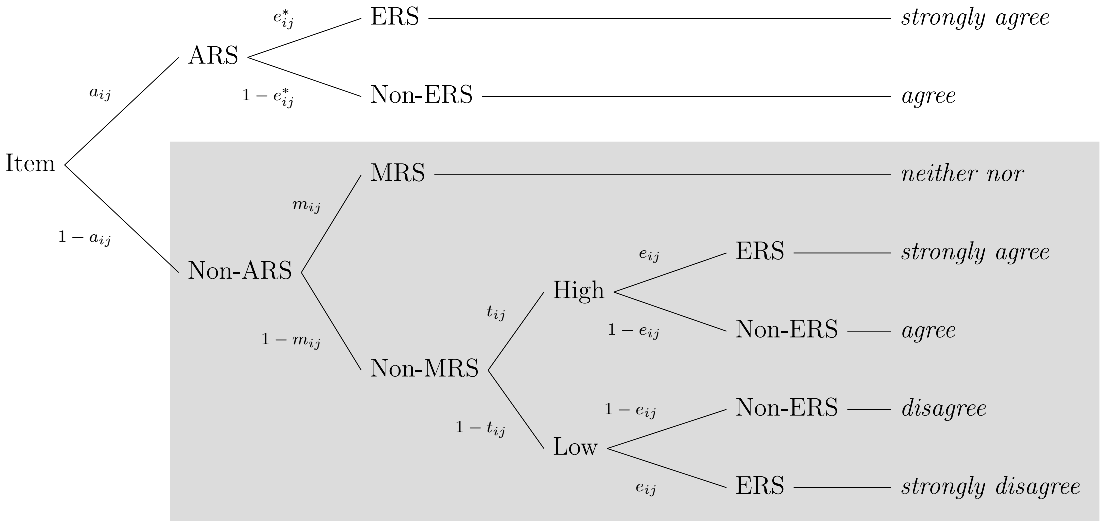


FIGURE 3.1: Tree diagram of the Acquiescence Model. The model includes the Böckenholt Model (depicted in gray) as a special case if $a_{ij} = 0$.

3.1 Model Development

MPT models assume, like IR-tree models, that a finite number or latent processes can explain the multinomial distribution of observed, categorical responses (Erdfelder et al., 2009; Hütter & Klauer, 2016; Riefer & Batchelder, 1988). In contrast to IR-tree models, multiple paths may lead to the same response category (as long as the model is identified), but standard MPT models do not incorporate person and/or item effects. This latter limitation, however, was relaxed in the recently developed class of hierarchical MPT models (Klauer, 2010; Matzke et al., 2015;

Smith & Batchelder, 2010). Therein, the MPT parameters are transformed using an appropriate link function (e.g., probit) and reparameterized, for example, using a person parameter θ_i and an item parameter β_j—just as in a standard IRT model. Thus, the model equation, for instance, for parameter m_{ij} is then $m_{ij} = \Phi(\theta_{mi} - \beta_{mj})$. That is, the probability of a midpoint response is higher the higher a person's MRS level θ_{mi} and the lower an item's MRS difficulty β_{mj}.

We built on these developments and showed that the Böckenholt Model can also be conceptualized as a hierarchical MPT model. Furthermore, this more general framework allowed us to develop the so-called *Acquiescence Model* depicted in Figure 3.1. Therein, an ARS branch is added to the Böckenholt Model such that affirmative responses are assumed to come from either a high target trait level or from ARS.

In the paper, we demonstrated in tailored simulation studies that the model parameters can be correctly recovered using the proposed Bayesian estimation procedure; and that it was possible to empirically discriminate between the Böckenholt Model and the Acquiescence Model. Furthermore, an empirical example from personality psychology was used to illustrate the interpretation of the model parameters, to assess the fit of the model also in comparison to other models, and to highlight that the Bayesian estimation framework can handle a model with nine correlated latent variables in a straightforward manner (which would almost be impossible using, for example, an expectation-maximization [EM] algorithm).

3.2 Acquiescence

In order to fully understand the implications of the proposed Acquiescence Model, it is instructive to compare it to other models in terms of the implied conception of ARS. A typical definition of ARS is, for example, given by Weijters, Geuens, and Schillewaert (2010b): "Respondents vary in their tendency to use positive response categories" (p. 96). This and other definitions, descriptions, and operationalizations conceptualize ARS in terms of what can be observed—namely, systematically more agreement than what would be expected on the basis of a person's target-trait level. However, these definitions do not describe and explain an underlying psychological process. All approaches enclose the possibility that ARS may lead to agreement when one would rather expect disagreement given the target-trait level. However, the definitions remain silent with respect to the following questions: Is ARS an ordinal process that may shift a 4 into a 5 or a 1 into a 2? Or, are disacquiescence and acquiescence two sides of the same coin, and may (low) ARS thus shift a 4 into a 2? Answering such questions would lead to a more precise description of ARS and enhance our understanding of this response style.

While the process of ARS is rarely if ever described in such detail, statistical models for ARS are of course more concrete. And it turns out that ARS is conceptualized as a shift process in the most prominent ARS models, which have been proposed in the framework of factor analysis, so-called bi-factor or random-intercept models (e.g., Billiet & McClendon, 2000; Ferrando et al., 2016; Kam & Zhou, 2015; Maydeu-Olivares & Coffman, 2006). That is, the following, very generic equation describes the relevant features of these models:

$$f(x_{ij}) = \lambda_j \theta^*_{ti} + \theta^*_{ai} - \beta_j.$$

Even though the models differ with respect to certain aspects of the equation, they share the notion that some target trait parameter θ^*_{ti} and some ARS parameter θ^*_{ai} act additively on the latent scale. Thus, ARS simply shifts the target trait up or down and is conceptualized as an ordinal process. Moreover, acquiescence and disacquiescence are then two sides of the same coin. In short, ARS may shift a 4 into a 5, a 1 into a 2, or a 4 into a 2—and this may or may not be congruent with one's definition of ARS. At least, if a sensible concept of ARS involves such predictions, we should strive for more general descriptions than something like "yeasaying" (e.g., Couch & Keniston, 1960).

Unlike shift models, the model proposed in the second paper takes a different route. Therein, agreement is a mixture of two components, namely, agreement stemming from the target trait and agreement stemming from ARS. As illustrated in Figure 3.1 and in more detail in the paper, ARS increases the probability of the two agree categories and simultaneously decreases the probability of the other three categories—and may of course change a 2 into a 4. But ARS may not change a 4 into a 5 or a 1 into a 2, this is solely influenced by the ERS process. Moreover, a very low level of ARS implies the absence of an ARS effect and not disacquiescence. That is, a symmetric response distribution is predicted in the absence of ARS (given, of course, intermediate item difficulty and target trait level). This model is in line with conceptions of ARS that emphasize the qualitative aspect of agreement and with researchers that use the number of agree responses as a measure of ARS (e.g., Billiet & McClendon, 2000). Furthermore, it is, contrarily to a shift model, in line with almost the only profound theoretical account of ARS proposed by Knowles and Condon (1999).

3.3 Discussion

In summary, we showed that the popular class of IR-tree models are special cases of the more general framework of hierarchical MPT models. Within this larger framework, it was possibly to extend an existing model for ERS and MRS to ARS. This development allows researchers interested in ARS to adopt the attractive, process-oriented perspective of IR-tree models. Moreover, it is now possible to compare the mixture approach to ARS with the established shift approach, and this is important for three reasons. First, a precise *description* is a prerequisite for scientific reasoning and research in general. Second, having a precise description of ARS helps to develop appropriate *explanations* of the process—and improvements in description and explanation are mutually reinforcing. Third, a deeper understanding of ARS will help to shed light on the commonalities and differences between ARS and, for example, other response styles, item-wording effects, careless responding, or socially desirable responding. With respect to a deeper understanding of ARS, the second paper answered some questions (e.g., how to incorporate ARS into IR-tree models), but it also raises some new questions (e.g., what is actually ARS exactly). But, science makes progress not only through the answers we give, but also through the questions we ask.

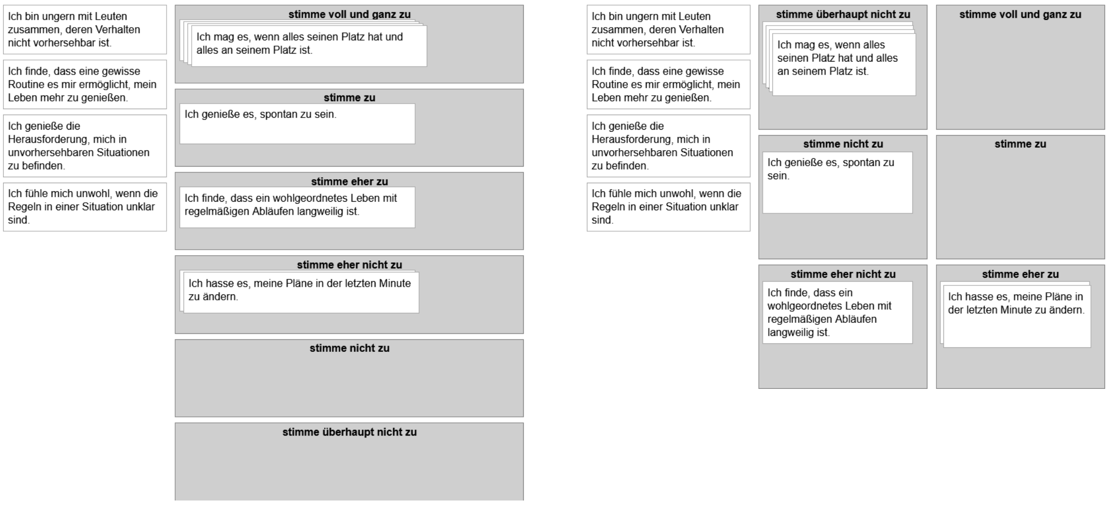
A further and also more ambitious aim of the second paper was to raise awareness for hierarchical MPT models among psychometricians. Even though the papers by Klauer (2010) and Matzke et al. (2015) were published in *Psychometrika*, they were focused on cognitive, experimental settings. However, hierarchical MPT modeling can potentially be a fruitful framework for future applications and developments in psychometrics beyond response styles. For example, it is well known that correctly answering a test item may often be the result of one of two (or more) processes: For example, both a visual and analytical strategy may help to correctly solve a cube-rotation item; or, a correct response may stem from knowledge but it may also stem from guessing, cheating, or previous item exposure. The framework of hierarchical MPT models may allow to develop new, tailored models for such situations or to re-think existing models. Thus, the second paper is a contribution with respect to IR-tree models and response styles, but also for psychometrics in general.

4 Towards a Better Response Format (and Back)

Plieninger, H., Henninger, M., & Meiser, T. (2017). *An experimental comparison of the effect of different response formats on response styles.* Manuscript in preparation.

If one wants to reduce the influence of response styles, one may try to implement control post hoc after data collection by means of an appropriate statistical method. However, it may be hard for one reason or another to implement this in day-to-day usage. For example, fitting the Acquiescence Model proposed in the second paper (Plieninger & Heck, 2017) may not be feasible in all situations. This leads to the question whether response styles can be controlled a priori. In the first paper (Plieninger, 2016), it was demonstrated that the use of reverse-coded items can be an effective means of ARS control. Apart from that, researchers have investigated different modifications of the Likert-type response format in order to further reduce response style variance. However, altering the anchoring labels or the number of categories—to give just one example—has not led to a resolution of the problem of response styles (e.g., Weijters, Cabooter, & Schillewaert, 2010). Thus, it might be the case that the solution lies outside the box, namely, in a different response format.

In a recent study, Böckenholt (2017) renewed an idea of Thurstone (1928) and proposed a drag-and-drop format, which is illustrated in Figure 4.1B. Therein, the respondent drags, with the computer mouse, each item from the left into the chosen category on the right. Böckenholt compared an IR-tree model and an ordinal graded response model across different formats and concluded: "The drag-and-drop method stands out because it triggered fewer response style effects than the other response formats. If this finding can be replicated in future research, one could argue that had Thurstone's (1928) approach been adopted instead of Likert's (1932) approach, response styles would play a much smaller role than they do now" (2017, p. 80).



Ich bin ungern mit Leuten zusammen, deren Verhalten nicht vorhersehbar ist.

Ich finde, dass eine gewisse Routine es mir ermöglicht, mein Leben mehr zu genießen.

Ich genieße die Herausforderung, mich in unvorhersehbaren Situationen zu befinden.

Ich fühle mich unwohl, wenn die Regeln in einer Situation unklar sind.

stimme voll und ganz zu

Ich mag es, wenn alles seinen Platz hat und alles an seinem Platz ist.

stimme zu

Ich genieße es, spontan zu sein.

stimme eher zu

Ich finde, dass ein wohlgeordnetes Leben mit regelmäßigen Abläufen langweilig ist.

stimme eher nicht zu

Ich hasse es, meine Pläne in der letzten Minute zu ändern.

stimme nicht zu

stimme überhaupt nicht zu

(A) Response format DnD-I.

Ich bin ungern mit Leuten zusammen, deren Verhalten nicht vorhersehbar ist.

Ich finde, dass eine gewisse Routine es mir ermöglicht, mein Leben mehr zu genießen.

Ich genieße die Herausforderung, mich in unvorhersehbaren Situationen zu befinden.

Ich fühle mich unwohl, wenn die Regeln in einer Situation unklar sind.

stimme überhaupt nicht zu

Ich mag es, wenn alles seinen Platz hat und alles an seinem Platz ist.

stimme voll und ganz zu

stimme nicht zu

Ich genieße es, spontan zu sein.

stimme zu

stimme eher nicht zu

Ich finde, dass ein wohlgeordnetes Leben mit regelmäßigen Abläufen langweilig ist.

stimme eher zu

Ich hasse es, meine Pläne in der letzten Minute zu ändern.

(B) Response format DnD-II.

FIGURE 4.1: Drag-and-drop formats used in Study 2 of Plieninger et al. (2017). To-be answered items appear on the left, and already answered items appear in the chosen category on the right.

4.1 Method and Results

In the third paper, we started with the hypothesis that the drag-and-drop format may indeed be an effective means of a-priori control of response styles as suggested by Böckenholt (2017). We conducted two experiments that were designed to shed light on the process(es) that may potentially lead to such an advantage of drag and drop over Likert. Our analyses included the comparison of an IR-tree model with an ordinal graded response model, a multi-group variant of the MPCM, as well as comparisons of reliability and validity across response formats. Our results revealed three interesting findings: First, the drag-and-drop format depicted in Figure 4.1B was less prone to response styles compared to the Likert-type format as revealed by the IR-tree model. This replicated the findings of Böckenholt (2017). Second, there was no such advantage in the three conditions that used a drag-and-drop format with only one column of response categories (see Figure 4.1A). Third, the differences between response formats across all conducted analyses was rather small. In summary, we found a small advantage of drag and drop only when the six categories were presented in two columns. Furthermore, the Likert-type condition performed at least as good if not better than drag and drop with only one column. Thus, we concluded that claiming a positive effect of the drag-and-drop format on response styles is premature if not even unwarranted.

Even though the drag-and-drop format may not resolve the problem of response styles, it may nevertheless be interesting to investigate in future research what has led to the advantage of the 2-column format. Our data revealed that respondents in this condition made more use of extreme categories but in a content-related and not style-related sense. This could potentially be caused by the more compact display of response categories (see Figure 4.1), which may have made the extreme categories more attractive. However, the gradual ordering of response categories is more explicit in the 1-column compared to the 2-column format, and this is usually desired in order to help respondents to interpret the meaning of the categories. Thus, identifying the psychological process(es) responsible for the advantage of the 2-column format remains a route for future research.

4.2 Likert-Type Response Format

Since the original paper by Likert (1932), the Likert-type format has been criticized, for instance, for fostering response styles (e.g., Brown & Maydeu-Olivares, 2012), for falsely implying an interval-scale nature of the data, and for numerous other abuses and misinterpretations (e.g., Carifio, Perla, Carifio, & Perla, 2007). However, alternative response formats such as the drag-and-drop format or a forced-choice format exhibit small to no advantages over the Likert-type format and/or are difficult to deal with for the researcher. Moreover, particular aspects related to the Likert-type format have been discussed in the literature in detail, for example, how many categories to use, how many and what anchors to use, or whether to present the categories in ascending and descending order (e.g., De-Castellarnau, 2017). But, all this research has hardly led to any groundbreaking insights or developments beyond good practices already known decades ago.

Thus, I take the opportunity to defend the Likert-type format on the basis of the literature reviewed herein and based on my own experience with Likert-type data as reported herein and beyond. First, the format is so heavily used that trying to replace it might be a waste of resources in the first place. Second, Likert-type items are easy to develop, answer, score, and analyze. Third, alternative formats such as a drag-and-drop format seem not to offer general and large benefits. Thus, I believe that we should put less effort in studies and discussions about the Likert-format itself, for example, about the optimal number of response categories. Rather, more effort should be put in developing reliable and valid items and scales and in a better understanding of the response process. In analogy to the criticized albeit ubiquitous α-level of .05: It seems impossible to abandon it, it's better to have an imperfect than no standard, and there are probably more important things (e.g., problem of underpowered studies) than $\alpha = .05$.

5 Discussion

5.1 Summary

This thesis has led to a deeper understanding of response styles. In short, I showed in the first paper (Plieninger, 2016) that response styles have severe detrimental effects only when correlated with the target trait. The new model for acquiescence proposed in the second paper (Plieninger & Heck, 2017) extended the scope of the popular class of IR-tree models. In the third paper (Plieninger et al., 2017), we empirically tested—and rejected—the hypothesis that a new response format, namely, a drag-and-drop format, may be able to solve the problem of response styles. In the following, the contributions will be integrated with respect to the psychometric, applied, and substantive level of response styles.

5.1.1 Psychometric Level

On the psychometric level, we showed that IR-tree models are special cases of hierarchical MPT models, we proposed a new model for ARS, and we employed and compared recent IRT models in various variants. In detail, a new model for ARS was proposed in the second paper. This model is built on recent advances in both psychometric and cognitive modeling, namely, IR-tree and hierarchical MPT models (Böckenholt, 2012; De Boeck & Partchev, 2012; Matzke et al., 2015). Our work is an example of the benefits that can emerge when models, techniques, or theories from different fields are brought together to solve problems where one field alone can reach only limited solutions. Furthermore, we made use of recent advances in Bayesian hierarchical modeling and respective software such as Stan (Carpenter et al., 2017) that allowed us to estimate this complex model comprised of up to nine latent variables.

Apart from that, psychometrics was also a recurring theme in the other two papers. In the first paper, the MPCM, an IRT model that was recently proposed by Wetzel and Carstensen (2017), was used as a data-generating model in simulation studies. The paper highlights the flexibility and usefulness of the model, and it shows that the model can be fit to empirical data also with software other than ConQuest—which was successfully used by Wetzel and Carstensen (2017)—namely, with the R package TAM (Robitzsch, Kiefer, & Wu, 2017). The MPCM was also used in the third paper, where an extension of the model allowed us to additionally include content-heterogeneous items in order to measure response styles more precisely.

With respect to psychometrics and quantitative methods more general, the present thesis illustrates the rich toolbox that is available to psychologists today. The conducted studies highlight how specific models and techniques can be selected and combined, not for their own sake, but rather to answer important questions in order to advance the understanding of the topic at hand. In all three papers, we used IRT models such as MPCMs (Wetzel & Carstensen, 2017), IR-tree models (Böckenholt, 2012; De Boeck & Partchev, 2012), or steps models (Tutz, 1990; Verhelst, Glas, & de Vries, 1997). Further methods and techniques—implemented in tailored software such as R, Stan, or Mplus—were employed where appropriate: For example, it was made use of MPT models, confirmatory factor analysis, simulation studies, Bayesian modeling and posterior predictive checking, or empirical analyses of both conducted experiments and existing data, to name a few.

5.1.2 Applied Level

On the applied level, we answered the question of bias caused by response styles, we evaluated a new response format, and we developed a new model that is ready to be applied in future work. More specifically, the simulation study in the first paper revealed that there is no need to fear a large detrimental impact of response styles in general. However, in situations were target trait and response style are substantially correlated, the situation changes and bias grows with increasing correlation. Apart from that, a new response format, drag and drop, was applied and evaluated in the third paper with a focus on response styles, reliability, and validity. The conducted experiments showed that the format leads to data roughly comparable to a Likert-type format and can thus be applied if desired—even though advantages over a Likert-type format are not to be expected. Last, future applications of our Acquiescence Model proposed in the second paper are facilitated through our R and Stan code that is publicly available.

5.1.3 Substantive Level

On the substantive level, we contributed to the understanding of response styles by delineating conditions under which response styles are most influential, by comparing qualitatively different accounts of ARS, and by pointing out open questions and future directions throughout this thesis. In more detail, the first paper clearly showed that special attention should be payed to correlates of response styles. Such correlates may be personality, motivation, cognitive capacity, or culture, but potentially also features of the situation or the item and questionnaire (e.g., Johnson, Shavitt, & Holbrook, 2011; Knowles & Condon, 1999; Krosnick, 1991; Schwarz, 1999; Shulruf, Hattie, & Dixon, 2008; Tourangeau & Rasinski, 1988). However, many open questions concerning such antecedents of response styles remain as pointed out in the Introduction above and in the literature (e.g., Van Vaerenbergh & Thomas, 2013; Wetzel, Böhnke, & Brown, 2016). Future research should not only aim to build a coherent nomological net around response styles, but should also pay careful attention to the causal structures between dependent variables, independent variables, and response styles as alluded to in the first paper. Apart from that, we pointed out in the second paper that two substantive interpretations of ARS exist, namely, either in terms of a shift or a mixture process. Our model brought up this previously overlooked question and made it possible to compare the two approaches.

5.2 Comparison of IRT Models for Response Styles

In all three papers, specific IRT Models for response styles were used, namely, IR-tree models and MPCMs. Many other models inside and outside of IRT exists, and overviews can be found in the literature (e.g., Böckenholt & Meiser, 2017; Henninger & Meiser, 2017; Wetzel, Böhnke, & Brown, 2016). In the following, the two model classes used herein will be compared. The MPCM is an extension of the traditional partial credit model in that additional latent variables for response styles are specified (Plieninger, 2016; Wetzel & Carstensen, 2017). Thus, the model retains the ordinal relationship between the responses and the target trait, and it reduces to an ordinal model when response style variance is zero. Similar approaches have been proposed in the literature (e.g., Falk & Cai, 2016; Jin & Wang, 2014; Johnson & Bolt, 2010). IR-tree models pursue a different route by assuming that a psychological meaningful tree-like structure—as, for instance, depicted in Figure 3.1—can explain the ordinal responses. Thus, the latent variables pertain to, in most cases, dichotomous decisions such as agreement vs. disagreement or extremity vs. modesty. Because of the complex tree structure, ordinal models do not exist as special cases for most IR-tree models; comparisons with non-nested ordinal models can either be performed with other IR-tree models such as a steps model (De Boeck & Partchev, 2012; Tutz, 1990) or other IRT models such as a graded response model (Böckenholt, 2017; Samejima, 1969).

Both model classes have in common that target traits and response styles are conceptualized as continuous latent variables. Moreover, they have similar concepts of specific response styles. For example, the weights for MRS used in an MPCM are $(0, 0, 1, 0, 0)$, and this is exactly the coding scheme for the first pseudoitem in an IR-tree model. However, important differences between the models exist as well: The most important one is probably the measurement of the target trait. In an MPCM, the target trait is measured using the ordinal information from the responses just as in a model without response styles. In an IR-tree model, in contrast, the target trait is only measured using binary information of agreement vs. disagreement. Even though Plieninger and Meiser (2014) showed that this did not impair the validity of the target trait, IR-tree models that take ordinal information into account (Meiser et al., 2017) are certainly a promising route for future research. A further difference concerns flexibility. While MPCMs are highly flexible and can accommodate different response styles for different numbers of categories (Falk & Cai, 2016; Plieninger, 2016; Wetzel & Carstensen, 2017), IR-tree models are less flexible, because structurally different models need to be specified for different numbers of categories (Böckenholt, 2012; De Boeck & Partchev, 2012; Plieninger & Meiser, 2014), and extensions, for example, to ARS are complex (Plieninger & Heck, 2017). In summary, the contribution of IR-tree models is that they focus on the psychological processes behind the responses and point to future research on the substantive level. The advantage of MPCMs is the retained ordinality and their flexibility, which makes them easier to be used on the applied level.

5.3 Future Directions

Open questions remain that this thesis could not address or brought up. First, it cannot be ignored that there are empirical examples that showed a bias due to response styles (e.g., Rammstedt, Goldberg, & Borg, 2010). It would be interesting to look at such data in detail in order to delineate whether the observed bias can be explained by findings from my simulations studies. Second, methods and models to control response styles need to be carefully compared (see Wetzel, Böhnke, & Rose, 2016, for an example); we need to know whether they are effective under a variety of conditions—especially when content and style are correlated—and whether they themselves might introduce other biases (like anchoring vignettes; von Davier, Shin, Khorramdel, & Stankov, 2017). Third, our proposed mixture model for acquiescence brought up the question what acquiescence really is. Future research should investigate whether the empirical phenomenon of acquiescence is better explained by a shift or a mixture account, or whether both are required. Furthermore, ARS should be compared to phenomena like item-wording effects, socially desirable responding, or careless responding in terms of the psychological processes involved. Such research may at some point also address the question whether—in terms of the underlying psychological processes—the directional response style of ARS is qualitatively different from other, symmetric response styles like ERS and MRS. Apart from that, hierarchical MPT models may help to solve psychometric problems outside the area of experimental psychology, and our research is an exemplar thereof. Finally, although we provided coherent evidence against a general advantage of a drag-and-drop format, future studies and replications may investigate the reported effect of the two-column format. In this context, our results indicated that responses in the 2-column condition were more variable in a beneficial way. It might be interesting to evaluate different ways (format, instructions, etc.) to achieve that same effect.

5.4 Conclusions

From my point of view, the major challenge to be addressed in future response style research is to advance the field on the substantive level. What is needed is a precise description of response styles and their psychological determinants. This would help, on the psychometric level, to compare existing and develop new models for response styles. And this would also help, on the applied level, to guide users whether and how response styles should be taken into account. However, as pointed out in the Introduction, this needs a shift (or possibly a mixture) of strategies: These goals won't be accomplished by simply publishing more applications and more models. What is needed is dedicated, persistent, and collaborative effort to accomplish these tasks, and I hope that the present thesis is a little piece of this puzzle.

6 References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi:10.1177/0146621697211001

Bachman, J. G. & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response style. *Public Opinion Quarterly*, *48*, 491–509. doi:10.1086/268845

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two-dimensional interpretation of acquiescence. *Psychological Bulletin*, *76*, 186–204. doi:10.1037/h0031474

Berg, I. A. & Collier, J. S. (1953). Personality and group differences in extreme response sets. *Educational and Psychological Measurement*, *13*, 164–169. doi:10.1177/001316445301300202

Billiet, J. B. & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628. doi:10.1207/S15328007SEM0704_5

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

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

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*, 159–181. doi:10.1111/bmsp.12086

Bolt, D. M. & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335–352. doi:10.1177/0146621608329891

Brown, A. & Maydeu-Olivares, A. (2012). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*, 36–52. doi:10.1037/a0030641

Carifio, J., Perla, R. J., Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*, 106–116. doi:10.3844/jssp.2007.106.116

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. doi:10.18637/jss.v076.i01

Cloud, J. & Vaughan, G. M. (1970). Using balanced scales to control acquiescence. *Sociometry*, *33*(2), 193–202. Retrieved from http://www.jstor.org/stable/2786329

Couch, A. & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*, 151–174. doi:10.1037/h0040372

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology*, *33*, 401–415. doi:10.1037/h0054677

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119–130. doi:10.1016/j.jrp.2015.05.004

De Boeck, P. & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. doi:10.18637/jss.v048.c01

DeCastellarnau, A. (2017). A classification of response scale characteristics that affect data quality: A literature review. *Quality & Quantity*. Advance online publication. doi:10.1007/s11135-017-0533-4

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, *217*, 108–124. doi:10.1027/0044-3409.217.3.108

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

Ferrando, P. J. & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 427–448. doi:10.1348/000711009X470740

Ferrando, P. J., Morales-Vives, F., & Lorenzo-Seva, U. (2016). Assessing and controlling acquiescent responding when acquiescence and content are related: A comprehensive factor-analytic approach. *Structural Equation Modeling*, *23*, 713–725. doi:10.1080/10705511.2016.1185723

Fischer, R. (2004). Standardization to account for cross-cultural response bias. *Journal of Cross-Cultural Psychology*, *35*, 263–282. doi:10.1177/0022022104264122

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125. doi:10.1177/014662169602000201

Henninger, M. & Meiser, T. (2017). *An integration of IRT models accounting for response styles*. Manuscript in preparation.

Hütter, M. & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, *27*, 116–159. doi:10.1080/10463283.2016.1212966

Jeon, M. & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085. doi:10.3758/s13428-015-0631-y

Jin, K.-Y. & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*, 116–138. doi:10.1177/0013164413498876

Johnson, T. R. & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*, 92–114. doi:10.3102/1076998609340529

Johnson, T. R., Shavitt, S., & Holbrook, A. L. (2011). Survey response styles across cultures. In D. Matsumoto, F. J. R. van de Vijver, U. Schönpflug, & E. van de Vliert (Eds.), *Cross-cultural research methods in psychology* (pp. 130–175). doi:10.1017/CBO9780511779381.008

Kam, C. C. S. & Zhou, M. (2015). Does acquiescence affect individual items consistently? *Educational and Psychological Measurement*, *75*, 764–784. doi:10.1177/0013164414560817

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*, 161–177. doi:10.1080/00273171.2013.866536

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98. doi:10.1007/s11336-009-9141-0

Knowles, E. S. & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*, 379–386. doi:10.1037/0022-3514.77.2.379

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*, 213–236. doi:10.1002/acp.2350050305

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 5–55.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*, 317–323. Retrieved from http://www.jstor.org/stable/24529203

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*, 205–235. doi:10.1007/s11336-013-9374-9

Maydeu-Olivares, A. & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344–362. doi:10.1037/1082-989X.11.4.344

Meiser, T. & Machunsky, M. (2008). The personal structure of personal need for structure: A mixture-distribution Rasch analysis. *European Journal of Psychological Assessment*, *24*, 27–34. doi:10.1027/1015-5759.24.1.27

Meiser, T., Plieninger, H., & Henninger, M. (2017). *Ordinal and multidimensional IRTree models for analyzing response styles and trait-based rating responses*. Manuscript in preparation.

Mirowsky, J. & Ross, C. E. (1991). Eliminating defense and agreement bias from measures of the sense of control: A 2 × 2 index. *Social Psychology Quarterly*, *54*, 127–145. doi:10.2307/2786931

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (Vol. 1, pp. 17–59). San Diego, CA: Academic Press.

Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32–53. doi:10.1177/0013164416636655

Plieninger, H. & Heck, D. W. (2017). *A new model for acquiescence at the interface of psychometrics and cognitive psychology*. Manuscript submitted for publication.

Plieninger, H., Henninger, M., & Meiser, T. (2017). *An experimental comparison of the effect of different response formats on response styles*. Manuscript in preparation.

Plieninger, H. & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899. doi:10.1177/0013164413514998

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality*, *44*, 53–61. doi:10.1016/j.jrp.2009.10.005

Ray, J. (1979). Is the acquiescent response style problem not so mythical after all? Some results from a successful balanced F scale. *Journal of Personality Assessment*, *43*, 638–643. doi:10.1207/s15327752jpa4306_14

Riefer, D. M. & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339. doi:10.1037/0033-295X.95.3.318

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules (Version 2.8-12). Retrieved from https://github.com/alexanderrobitzsch/TAM

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, *63*, 129–156. doi:10.1037/h0021888

Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann.

Rust, J. (2009). *Modern psychometrics: The science of psychological assessment* (3rd ed.). London, UK: Routledge.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Society. Richmond, VA. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Savalei, V. & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*, 407–424. doi:10.1080/00273171.2014.931800

Schimmack, U., Böckenholt, U., & Reisenzein, R. (2002). Response styles in affect ratings: Making a mountain out of a molehill. *Journal of Personality Assessment*, *78*, 461–483. doi:10.1207/S15327752JPA7803_06

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93–105. doi:10.1037/0003-066X.54.2.93

Shulruf, B., Hattie, J., & Dixon, R. (2008). Factors affecting responses to Likert type questionnaires: Introduction of the ImpExp, a new comprehensive model. *Social Psychology of Education*, *11*, 59–78. doi:10.1007/s11218-007-9035-x

Smith, J. B. & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. doi:10.1016/j.jmp.2009.06.007

ten Berge, J. M. F. (1999). A legitimate case of component analysis of ipsative measures, and partialling the mean as an alternative to ipsatization. *Multivariate Behavioral Research*, *34*, 89–102. doi:10.1207/s15327906mbr3401_4

Thissen-Roe, A. & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*, 522–547. doi:10.3102/1076998613481500

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554. doi:10.1086/214483

Tourangeau, R. & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *103*, 299–314. doi:10.1037/0033-2909.103.3.299

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55. doi:10.1111/j.2044-8317.1990.tb00925.x

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*, 195–217. doi:10.1093/ijpor/eds021

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). doi:10.1007/978-1-4757-2691-6_7

von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2017). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement*. Advance online publication. doi:10.1177/0146621617730389

Wang, W.-C., Wilson, M., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement*, *43*, 335–353. doi:10.2307/20461834

Weijters, B., Cabooter, E. F. K., & 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*, 236–247. doi:10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*, 105–121. doi:10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*, 96–110. doi:10.1037/a0018721

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC*

International Handbook of Testing and Assessment (pp. 349–363). doi:10.1093/med:psych/9780199356942.003.0024

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*, 304–324. doi:10.1177/0013164415591848

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

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178–189. doi:10.1016/j.jrp.2012.10.010

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment*, *23*, 279–291. doi:10.1177/1073191115583714

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports. *Journal of Personality*, *84*, 461–472. doi:10.1111/jopy.12172

A Co-Authors' Statements

Daniel W. Heck

It is hereby confirmed that the following paper included in the present thesis was primarily conceived and written by its first and main author Hansjörg Plieninger.

Plieninger, H. & Heck, D. W. (2017). *A new model for acquiescence at the interface of psychometrics and cognitive psychology.* Manuscript submitted for publication.

Daniel and Hansjörg developed together the idea of the Acquiescence Model, and Hansjörg contributed his knowledge of item response modeling and response styles. Hansjörg planned and carried out all simulation studies as well as the analyses of empirical data, and this involved also the development of large parts of the R and Stan code. Hansjörg was solely responsible for writing the first draft as was well as for the finalization and submission of the paper.

Daniel contributed his knowledge of MPT modeling and Bayesian statistics to the development of the Acquiescence model. Furthermore, he wrote the inital R-, Stan-, and JAGS-code snippets, and he gave helpful comments on draft versions of the paper. Apart from that, Daniel helped in numerous discussions to refine specific parts of the idea and studies.

Daniel W. Heck

Place, Date

Mirka Henninger

It is hereby confirmed that the following paper included in the present thesis was primarily conceived and written by its first and main author Hansjörg Plieninger.

Plieninger, H., Henninger, M., & Meiser, T. (2017). *An experimental comparison of the effect of different response formats on response styles.* Manuscript in preparation.

Hansjörg developed the idea of the paper as well as the design and procedure of the experiments, and he contributed to the collection of the data. He was solely responsible for all analyses, for writing the first draft and for the finalization of the paper.

Mirka contributed in numerous discussions to refine the design and procedure of the experiments and gave helpful comments on draft versions of the paper. Furthermore, she contributed to the collection of the data.

Mirka Henninger	Place, Date

Thorsten Meiser

It is hereby confirmed that the following paper included in the present thesis was primarily conceived and written by its first and main author Hansjörg Plieninger.

Plieninger, H., Henninger, M., & Meiser, T. (2017). *An experimental comparison of the effect of different response formats on response styles.* Manuscript in preparation.

Hansjörg developed the idea of the paper as well as the design and procedure of the experiments, and he contributed to the collection of the data. He was solely responsible for all analyses, for writing the first draft and for the finalization of the paper.

Thorsten contributed in discussions to refine the design and procedure of the experiments and gave helpful comments on draft versions of the paper.

Thorsten Meiser

Place, Date

Mountain or Molehill? A Simulation Study on the Impact of Response Styles

Hansjörg Plieninger
University of Mannheim

Abstract

Even though there is an increasing interest in response styles, the field lacks a systematic investigation of the bias that response styles potentially cause. Therefore, a simulation was carried out to study this phenomenon with a focus on applied settings (reliability, validity, scale scores). The influence of acquiescence and extreme response style was investigated, and independent variables were, for example, the number of reverse-keyed items. Data were generated from a multidimensional IRT model. The results indicated that response styles may bias findings based on self-report data, and that this bias may be substantial if the attribute of interest is correlated with response style. However, in the absence of such correlations, bias was generally very small, especially for extreme response style and if acquiescence was controlled for by reverse-keyed items. An empirical example was used to illustrate and validate the simulation. In summary, it is concluded that the threat of response styles may be smaller than feared.

This is the accepted version of the following article: Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32–53. doi:10.1177/0013164416636655. Copyright © The Author(s) 2016. Reprinted by permission of SAGE Publications. The article has been published in final form at SAGE. The R code for this paper may be retrieved from https://osf.io/kunf3/.

Introduction

There exists the widespread claim and fear that response styles—such as acquiescence response style (ARS) or extreme response style (ERS)—distort results based on self-report data. The goal of the present simulation study was to present data rather than claims and to scrutinize the effect of response styles. The study covered three scenarios of a prototypical psychological research process, namely, estimating the reliability of a scale, testing its validity via correlations, and assigning a score to every respondent. To closely mirror situations in the applied field, the simulated data were analyzed using basic procedures (e.g., Cronbach's alpha) without trying to control for response styles. The data generating model, however, was a rather complex item response model that allowed to flexibly cover a variety of conditions.

Response Styles

Response styles are defined as the tendency to respond to questionnaire items irrespective of content (cf. Nunnally, 1978). This does not imply that the subject matter is irrelevant to the respondent, but indicates that response styles act independently of content and that both sources influence the actual response. This theoretical notion is supported by empirical evidence showing that response styles are stable across content domains (e.g., Weijters, Geuens, & Schillewaert, 2010a; Wetzel, Carstensen, & Böhnke, 2013). Moreover, it is well documented that response styles are stable within a questionnaire as well as across periods of several years (e.g., Aichholzer, 2013; Weijters, Geuens, & Schillewaert, 2010b).

Response styles represent a source of interindividual variance—additional to the content-related variance—that is usually not taken into account in analyses of self-report data, at least in more applied settings. There seem to be three different viewpoints on the matter. First, probably the majority of practitioners and researchers ignore response styles, because they don't know enough about them or cannot implement (statistical) control for one reason or another. Second, some take the position that response styles are negligible because this source of variance is small, represents error variance, or is trifling compared to content (e.g., Rorer, 1965; Schimmack, Böckenholt, & Reisenzein, 2002). Third, many researchers believe that response styles are a serious threat to the quality of self-report data that potentially influence all kinds of measures scientists usually draw conclusions from. For example, Eid and Rauber (2000) stated that "differences in category use can distort the results [. . .]" (p. 21). Likewise, Weijters, Geuens, and Schillewaert (2010b) wrote that "response styles have been found to bias estimates of means, variances, and correlations [. . .], leading to potentially erroneous results and conclusions [. . .]" (p. 96).

Although individual findings support the impression that response styles form a severe threat, the literature lacks a systematic investigation of the amount of bias and the conditions under which bias occurs. Simulation studies are well suited to address this issue, because they allow a comprehensive analysis of a specific effect (e.g., of response styles) while having full control over all other influences. However, there are only very few studies published that attempt to look at response styles from the perspective of a simulation study. Heide and Grønhaug (1992) published a simulation in a marketing journal and found biasing effects of ARS and ERS, but their methodological approach was rather basic from today's perspective. The paper of Ferrando and Lorenzo-Seva (2010) also contains a simulation study on ARS with a limited range of conditions finding that ARS can bias results, but that this bias is minor for most practical purposes, at least with fully balanced scales (i.e., equal number of regular and reverse-keyed items). Savalei and Falk (2014) found that substantive factor loadings were only affected by ARS when its influence was strong. Wetzel, Böhnke, and Rose (2016) investigated trait recovery of different methods, which aim to control for ERS, and stated: "The results of our simulation study imply that ignoring ERS on average hardly affects trait estimates if ERS and the latent trait are uncorrelated or only weakly correlated [. . .]" (p. 17).

Statistical Models for Response Styles

A multitude of models to measure and/or control for response styles have been proposed, which vary greatly in terms of their objectives, requirements, and complexity (cf. Van Vaerenbergh & Thomas, 2013). For example, in confirmatory factor analysis, an additional acquiescence factor can be used to analyze scales comprised of both regular and reverse-keyed items (e.g., Billiet & McClendon, 2000). Different routes have been pursued in the family of item response theory (IRT). For example, mixture distribution Rasch models have been applied with the result that a 2-class solution could be interpreted as comprising non-extreme and extreme respondents (e.g., Eid & Rauber, 2000; Meiser & Machunsky, 2008; Wetzel et al., 2013). Böckenholt (2012) proposed a multidimensional IRT model in which the original response is separated into content- and response style-related processes using dichotomous pseudoitems (cf. De Boeck & Partchev, 2012; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014). Another multidimensional IRT model, namely, a variant of Bock's *nominal response model*, was developed by Bolt and colleagues (Bolt & Newton, 2011; Johnson & Bolt, 2010) and further extended by Falk and Cai (2016). Furthermore, multidimensionality arising from random thresholds is accounted for in models suggested by Wang (e.g., Jin & Wang, 2014).

Most of the models proposed so far focus on only one response style and cannot be modified to accommodate another one. However, Wetzel and Carstensen (2015) recently proposed an approach in the framework of multidimensional Rasch models that allows to take into account both ARS and ERS.

Multidimensional Rasch Models

Multidimensional Rasch models date back to Georg Rasch (1961) himself and have, since then, been presented in multiple ways. Herein, the notation of Adams, Wilson, and Wang (1997), who call their approach *multidimensional random coefficients multinomial logit model*, is adapted. Therein, it is assumed that—possibly multiple—latent variables drive the item responses in an additive manner. The model has only one type of item parameter, namely, a difficulty parameter, which herein—for the sake of simplicity—was parametrized using a *rating scale model* approach (cf. Andrich, 1978), but other versions of the model for ordinal and binary items exist. In the current study, it is furthermore assumed that a symmetric, bipolar response format is used (e.g., ranging from *strongly disagree* to *strongly agree*).

Assume we have item i $(i = 1, \ldots, I)$ with $K + 1$ response categories $(k = 0, 1, \ldots, K)$ and person j $(j = 1, \ldots, J)$. The model has d $(d = 1, \ldots, D)$ latent dimensions and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)'$ is a column vector containing one person parameter per dimension. In the rating scale model, the item parameters comprise item location parameters β_i reflecting the overall difficulty of an item and threshold parameters τ_k, which are constant across items. This results in $I + K$ different item parameters overall contained in the vector $\boldsymbol{\xi} = (\beta_1, \ldots, \beta_I, \tau_1 \ldots, \tau_K)'$. The threshold parameters are constrained to sum to zero, $\sum_1^K \tau_k = 0$[1]. If the model parameters are to be estimated from empirical data, additional restrictions on the person or on the item parameters have to be made, because the model is otherwise not identified (cf. Adams et al., 1997).

Both the item parameters and the person parameters are mapped onto the category probabilities using a *design matrix* $\mathbf{A}$ and a *scoring matrix* $\mathbf{B}$, respectively. The linear combination of item parameters pertaining to category k of item i is defined by a row vector $\mathbf{a}_{ik}$ (of length $I + K$). The matrix $\mathbf{A}_i$ comprises $K + 1$ of these row vectors stacked below each other and defines the design matrix for item i, and I of these matrices are then again stacked below each other defining

[1] All K τ parameters are explicitly displayed in the example below for consistency with the simulation set-up even though the constraint makes one of the τ parameters redundant.

the design matrix **A**. An example for two items with three categories is depicted below:

$$\mathbf{A} * \boldsymbol{\xi} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 2 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 2 & 1 & 1 \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \tau_1 \\ \tau_2 \end{bmatrix}.$$

The weight of category k of item i on each of the dimensions is defined by the row vector $\mathbf{b}_{ik}$ (of length D). The matrix $\mathbf{B}_i$ comprises $K+1$ of these row vectors stacked below each other and defines the design matrix for item i, and I of these matrices are then again stacked below each other defining the design matrix **B**. Three examples of scoring matrices for two items with three categories are depicted below. The first one is typically employed in polytomous, unidimensional models like the rating scale model. The second is an example of *between-item multidimensionality*, where each item loads on only one dimension. The last one is an example of *within-item multidimensionality*, where the second item loads on both dimensions:

$$\mathbf{B}^{(1)} * \theta = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 1 \\ 2 \end{bmatrix} * \begin{bmatrix} \theta_1 \end{bmatrix}; \quad \mathbf{B}^{(2)} * \boldsymbol{\theta} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 2 \end{bmatrix} * \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}; \quad \mathbf{B}^{(3)} * \boldsymbol{\theta} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \end{bmatrix} * \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}.$$

Then, the probability of a response falling in category k of item i is modeled as

$$P(X_{ik} = 1; \mathbf{A}, \mathbf{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) = \frac{\exp(\mathbf{b}_{ik}\boldsymbol{\theta} - \mathbf{a}_{ik}\boldsymbol{\xi})}{\sum_{k=0}^{K} \exp(\mathbf{b}_{ik}\boldsymbol{\theta} - \mathbf{a}_{ik}\boldsymbol{\xi})},$$

where $\mathbf{a}_{ik}$ and $\mathbf{b}_{ik}$, respectively, represent a row vector of **A** and **B**, respectively, pertaining to the kth category of item i. The model reduces to Andrich's rating scale model for $D = 1$ and to the Rasch model for $K = 1$ and $D = 1$.

Multidimensional Rasch Models for Response Styles

Previously, multidimensionality within items has been investigated in situations where items measure more than one dimension at a time (cf. Adams et al., 1997). Wetzel and Carstensen (2015) extended the idea of within-item multidimensionality noting that not all of the latent dimensions need necessarily be related to the content of the items, but could also be related to, for example, response styles. This, in turn, requires different weights composing the matrix **B**. Assuming that each response involves one attribute- and one response style-dimension, scoring matrices for an item with five categories involving ERS and ARS, respectively, may look as follows (Wetzel & Carstensen, 2015):

$$\mathbf{B}^{(ERS)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 4 & 1 \end{bmatrix}; \quad \mathbf{B}^{(ARS)} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 3 & 1 \\ 4 & 1 \end{bmatrix}.$$

For ERS, the direction (agreement vs. disagreement) of the response is still governed by the first, content-related dimension alone; however, the extremity of the response may be altered by ERS. Contrarily, ARS may alter the direction of the response and may lead to, for example, agreement with both regular and reverse-keyed items.

In summary, multidimensional Rasch models are an interesting alternative to existing response style models. First, the model is very flexible: Various forms of response styles can be implemented; the only restriction is to find sensible weights for the matrix **B**. Second, this allows to simulate ERS and ARS from the same model facilitating the design of the study as well as the interpretation and comparison of the results. Third, multiple attribute-dimensions can be included. Fourth, the framework incorporates a unidimensional (content-only) model as a special case. Fifth, the model is parsimonious, because, for example, the number of item parameters is independent of the number of dimensions. These features make the model well-suited for the purposes of the present study, which aimed to investigate both ARS and ERS and which was intended to realistically cover situations of applied data analysis. Apart from that, even though it is a new model, the underlying notion of response styles is highly similar to that of established approaches (e.g., Johnson & Bolt, 2010; Weijters, Cabooter, & Schillewaert, 2010).

The Present Research

The present simulation study aimed to scrutinize the claim that response styles threaten the results of self-report data, especially so in applied settings. In more detail, the idea was to simulate data using the framework introduced above, to subsequently ignore response styles during data analysis (as is often done in the field), and finally to quantify the bias introduced by ARS or ERS. In order to cover a variety of settings, three different scenarios resembling prototypical steps of a research process were designed. First, Cronbach's alpha is arguably the most prominent measure of the reliability of a set of items, and it was investigated whether response styles would bias this measure (and how much). Second, the validity of a scale is often assessed using the correlation of two scale scores, and it was again investigated whether response styles would bias this measure (and how much). Third, the ultimate goal of assessment is to assign a score to every person. The accuracy of this was investigated (a) using correlations of true and observed scores and (b) by comparing the rank order of persons with and without response styles. In other words, it was examined how response styles may influence a decision (e.g., in health, education, work) that is based on self-report data. The analyses in all three scenarios employed raw score-based measures derived from classical test theory—for the reason that those measures are heavily used in applied research. This simulation study went beyond previous work in that the influence of both ERS and ARS was investigated under a wide range of conditions. Furthermore, the results of the simulation were verified and illustrated with an empirical example.

Method

Simulation Design and Set-Up

The simulation model had D dimensions comprising the attribute(s) of interest, θ_1 and possibly θ_2, (e.g., personality trait, attitude, symptom) and the response style θ_{RS}. In the case of two attributes, θ_1 and θ_2 each influenced a unique set of items (between-item multidimensionality), whilst θ_{RS} always influenced all items (within-item multidimensionality). In each replication, the person parameters were sampled from a multivariate normal distribution, $\boldsymbol{\theta} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \\ \mu_{\mathrm{RS}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{1,2} & \rho_{1,\mathrm{RS}} \\ \rho_{1,2} & \sigma_2^2 & \rho_{2,\mathrm{RS}} \\ \rho_{1,\mathrm{RS}} & \rho_{2,\mathrm{RS}} & \sigma_{\mathrm{RS}}^2 \end{bmatrix}.$$

If $\theta_{\mathrm{RS}} \sim N(0, 0)$, the response style dimension effectively drops making the model a content-only model.

In order to manipulate the amount of response style variance relative to substantive variance, σ_1^2 $(= \sigma_2^2)$ was fixed to a value of one. The amount of response style variance σ_{RS}^2 was varied between values of 0.00 and 1.00 (in steps of 0.10) indicating how diverse a sample is with respect to response styles, and higher values indicate more diversity. In each replication, the off-diagonal elements in $\mathbf{\Sigma}$ were drawn from a Wishart distribution with an identity matrix used as the *scale matrix* and $df = 10$; this results in the fact that the correlations have an expected value of zero and a variance of .10. The center of the response style distribution μ_{RS} was varied between values of -1.00 and 1.00 (in steps of 0.10). Positive (negative) values indicate that the sample overall tends to give more extreme responses (more non-extreme responses) for ERS and more (less) agree-responses for ARS, respectively.

Each replication entailed 200 persons and 10 items per attribute.[2] The number of categories was not varied and set to five (but see the Appendix). The number of reverse-keyed items was varied between zero and five per attribute. In each replication, the item location parameters β_i were drawn from a truncated normal distribution, TN(0, 1, -1.5, 1.5). The item threshold parameters τ_k were each drawn from a uniform distribution, U(-2.5, 2.5), and they were sorted in ascending order to avoid category reversals.[3] Subsequently, the thresholds were centered because of the restriction $\sum_1^K \tau_k = 0$ (and it was made sure that none of the τ parameters exceeded the limits of ±2.5). To illustrate the effect response styles have in the present model with the given set-up, an example is shown in Figure 1; data were generated for 100,000 people and an item of intermediate difficulty with equally spaced threshold parameters between -1.5 and 1.5. The figure shows, for example, that the uppermost third of the ERS distribution used the extreme categories twice as much compared to the baseline condition without response styles.

The scoring matrix $\mathbf{B}$ of each simulation model was adopted from the following template according to the number of attributes, their assigned number of regular and reverse-keyed items, and the type of response style:

$$\mathbf{B}' = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix} \begin{array}{l} \text{(regular item)} \\ \text{(reverse-keyed item)} \\ \text{(ERS)} \\ \text{(ARS).} \end{array}$$

Dependent Variables

In the first scenario, Cronbach's alpha was used as an estimate of the reliability of a set of items. To compare this value to a response style-free measure, it was made

[2] Pilot simulations revealed that N and I had virtually no effect on bias when varied between 100 and 1,000 and between five and 15, respectively.

[3] In the case of two attributes, the threshold parameters were equal for both attributes.

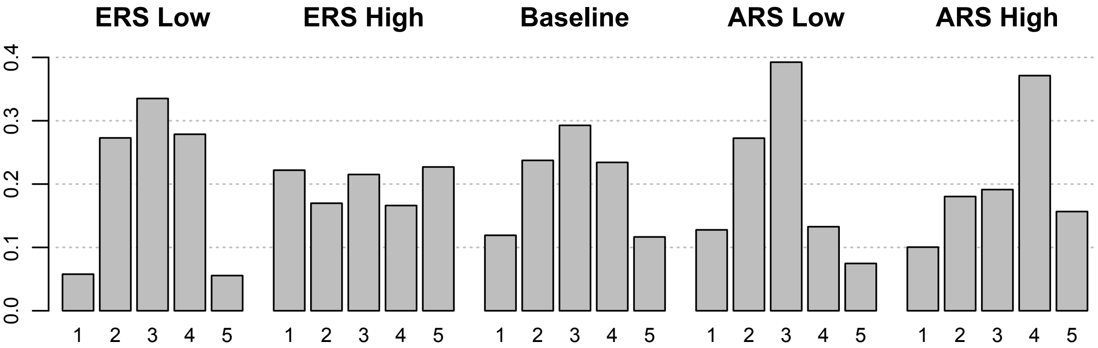


FIGURE 1: Illustration of the effect of response styles ($\mu_{\text{RS}} = 0$; $\sigma^2_{\text{RS}} = 1$; $\rho_{1,\text{RS}} = 0$). Displayed are responses to a 5-point item of the lower and the upper third of the—ERS or ARS—distribution as well as a baseline condition without response styles.

use of the concept of *covariate-free reliability* recently introduced by Peter Bentler (2016). He proposed a measure of *covariate-free alpha*, which controls Cronbach's alpha for the influence of a covariate (i.e., response styles in the present case) via partialing.[4] The actual dependent variable that was used in the analyses was the amount of bias, that is, the difference between Cronbach's alpha and covariate-free alpha.

In the second scenario, a scale score $\bar{x}_d$ (i.e., the mean across items after recoding) was computed for both attributes. The correlation of these two scale scores was compared to the partial correlation that controls the correlation of interest for response style ($r_{\bar{x}_1,\bar{x}_2\cdot\theta_{\text{RS}}}$). Again, bias was used as the dependent variable, that is, the difference between the observed and the partial correlation.

In the last scenario, the true person parameters θ_1, which are independent of response styles, were compared with the observed scale scores, which are influenced by both the attribute and response styles. First, these two variables were correlated. Second, the rank order of persons was compared at different cutoffs. For example, at a cutoff value of $c = .80$, people at or above the 80th percentile of scale scores were classified as *positive*. Additionally, this classification was done using the true person parameters θ_1 and the same cutoff c resulting in four possible outcomes: *true positives* (TP; originally and observed positive), *false positives* (FP; originally negative but observed positive), *false negatives* (FN; originally positive but observed negative), and *true negatives* (TN; originally and observed negative). To illustrate the results, three different measures at 39 equally spaced cutoffs between $c = .025$ and $c = .975$ were calculated in every replication: the *true positive*

[4] Regressing a true score T on a covariate Z yields a covariate-dependent part $T^{(Z)}$ and an orthogonal, covariate-free part $T^{\perp Z}$. Thus, it follows that $\sigma^2_T = \sigma^2_{T^{(Z)}} + \sigma^2_{T^{\perp Z}}$. Bentler (2016) showed that this holds also for the mean of the item-covariances (i.e., $\bar{\sigma}_{ij} = \bar{\sigma}^{(Z)}_{ij} + \bar{\sigma}^{\perp Z}_{ij}$), which is used in the equation of Cronbach's alpha, and proposed to decompose Cronbach's alpha into a covariate-dependent and a covariate-free part (i.e., $\alpha = \alpha^{(Z)} + \alpha^{\perp Z}$).

rate (TPR = TP/[TP + FN]) or *sensitivity* indicating how many of the people originally above the cutoff were indeed selected, the *false positive rate* (FPR = FP/[FP + TN]) indicating how many of the people originally below the cutoff were falsely selected, and the overall *accuracy* (ACC = [TP + TN]/[TP + FP + FN + TN]) indicating the total rate of correct classifications.

Results

The results are based on 100,000 replications for each simulation. In each single replication, the values of all variables were randomly and independently drawn. The simulations and analyses were conducted in R 3.2.1 (R Core Team, 2014).[5]

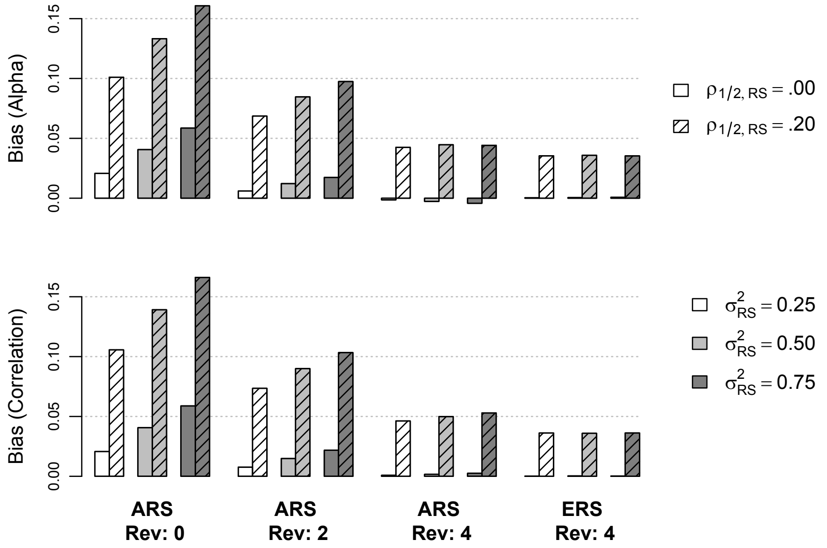


FIGURE 2: Overview of average bias with respect to Cronbach's alpha (upper panel) and the correlation of two scale scores (lower panel). Results are based on 1,000 replications in each of the selected conditions. $\rho_{1/2,\,\mathrm{RS}}$ stands for the correlation of the attribute (alpha) or each of both attributes (correlation) with the response style.

An overview of the average bias with respect to Cronbach's alpha (upper panel) and a correlation coefficient (lower panel) for selected conditions is given in Figure 2: ARS led to more bias compared to ERS, more reverse-coded items reduced bias, and more response style variance led to more bias. Furthermore, bias rarely exceeded levels of .05 if the attribute(s) and the response style were uncorrelated, but the opposite was true if the attribute(s) and the response style were moderately correlated. This figure gives already instructive insights, and more detailed results are reported in the following sections. In line with recommendations, for example, by Harwell (e.g., Harwell, Stone, Hsu, & Kirisci, 1996), it was chosen

[5] It was made use of the packages **MASS** (Venables & Ripley, 2002), **truncnorm** (Trautmann, Steuer, Mersmann, & Bornkamp, 2014), and **magic** (Hankin, 2005).

to refrain from presenting full-page tables with descriptive results. Rather, the results of each simulation were submitted to a regression model, which facilitates interpretation and makes it easier to detect effects of higher order. Unstandardized (b) and standardized (b^*) regression coefficients are reported.

Estimating the Reliability of a Scale in the Presence of Response Styles

Acquiescence

Two regression models were fit to the simulation results, one without and one with higher-order terms (see Table 1), and the following interpretation is based on the correctly specified, second model. Overall, the intercept indicated that—on average—the estimated alpha coefficient (which was .88) slightly overestimated the reliability by .01. Bias increased when fewer reverse-keyed items were used and when ARS variance was higher. Moreover, the substantive linear and quadratic effects of $\rho_{1,\,\text{ARS}}$ indicated that bias was most pronounced if ARS was related to the attribute of interest. Furthermore, an interaction effect indicated that reverse-keyed items buffer against the biasing effect of the attribute-ARS correlation. Both the interaction and the quadratic effect are illustrated in Figure 3 (left panel). There was no effect of μ_{RS} in this or any of the other simulations, because this parameter simply causes a shift of all responses without an effect on individual differences.

Table 1: Effect of Response Styles on Cronbach's Alpha as a Function of Reverse-Keyed Items and Joint Distribution of Attribute and Response Style

	ARS				ERS			
	Model 1		Model 2		Model 1		Model 2	
	b	b^*	b	b^*	b	b^*	b	b^*
Intercept	0.082		0.010		0.079		0.001	
Reversed	−0.006	−0.09	−0.006	−0.10	0.000	0.00	0.000	0.00
μ_{RS}	0.001	0.00	0.000	0.00	−0.001	−0.01	−0.001	−0.01
σ^2_{RS}	0.047	0.13	0.010	0.03	0.042	0.12	0.000	0.00
$\rho_{1,\,\text{RS}}$	0.139	0.37	0.140	0.38	−0.001	0.00	0.001	0.00
$(\rho_{1,\,\text{RS}})^2$			0.792	0.65			0.858	0.74
Reversed × $\rho_{1,\,\text{RS}}$			−0.055	−0.25				
R^2		0.17		0.95		0.02		0.96

Note. All predictor variables were centered. All SEs for paramters $b \leq .001$.

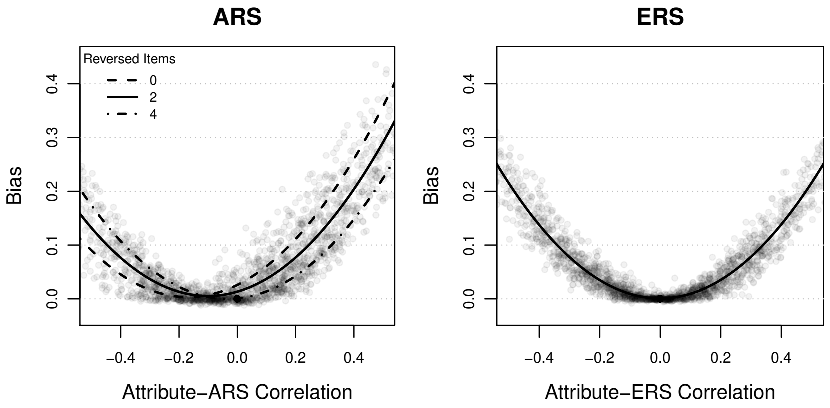


FIGURE 3: Effect of ARS and ERS, respectively, on Cronbach's alpha. Plotting region was restricted to $|\rho_{1,\mathrm{RS}}| < .5$ and a subset of 2,000 replications.

Extreme Response Style

This simulation focused on the effect of ERS on Cronbach's alpha. The intercept was virtually zero (Model 2) indicating that Cronbach's alpha was almost unbiased if $\rho_{1,\mathrm{ERS}} = 0$ (see Table 1). However, there was again a substantial quadratic effect of the attribute-ERS relationship, which is illustrated in Figure 3 (right panel). When the attribute and ERS were positively related, persons with a high (low) attribute level tend to give more (less) extreme answers. Thus, these responses undergo an upward-shift resulting in the fact that the items share additional variance that is due to ERS. When the relationship is negative, the responses are shifted downwards, which also increases the shared variance. This additional variance is wrongly attributed to the attribute if ERS is ignored leading to the observed bias.

Estimating the Correlation of Two Scales in the Presence of Response Styles

In addition to the previous scenario, the simulations now entailed further independent variables, namely, the correlation of the two attributes ($\rho_{1,2}$) as well as the correlation of the attributes with response style ($\rho_{1,\mathrm{RS}}$ and $\rho_{2,\mathrm{RS}}$, respectively).

Acquiescence

The results in Table 2 indicated that the actual correlation was, on average, slightly overestimated by a value of .01 when acquiescence was ignored as indicated by the intercept. Mirroring the results presented above, this bias became larger when fewer reverse-keyed items where employed and when ARS variance increased. Again, the center of the ARS distribution had no impact. Additionally, the negative slope of the true correlation between the two attributes indicated that bias became smaller the more positive the true relationship became. This is due to the fact that ARS makes correlations more positive, and the impact of this decreases the more positive the true correlation of the attributes already is. Note that this effect is only interpretable in the correctly specified, second model (see Table 2).

TABLE 2: Effect of Response Styles on Scale Score Correlation as a Function of Reverse-Keyed Items and Joint Distribution of Attributes and Response Style

	ARS				ERS			
	Model 1		Model 2		Model 1		Model 2	
	b	b^*	b	b^*	b	b^*	b	b^*
Intercept	0.012		0.012		0.000		0.000	
Reversed	−0.007	−0.12	−0.006	−0.11	0.000	0.00	0.000	0.00
σ^2_{RS}	0.024	0.08	0.024	0.08	0.000	0.00	0.000	0.00
μ_{RS}	0.002	0.01	0.002	0.01	0.000	0.00	0.000	0.00
$\rho_{1,2}$	0.005	0.02	−0.076	−0.25	0.015	0.05	−0.069	−0.24
$\rho_{1,\mathrm{RS}}$	0.079	0.25	0.082	0.26	−0.001	0.00	0.000	0.00
$\rho_{2,\mathrm{RS}}$	0.081	0.25	0.085	0.27	−0.001	0.00	0.001	0.00
$\rho_{1,\mathrm{RS}} \times \rho_{2,\mathrm{RS}}$			0.877	0.83			0.924	0.94
Reversed × σ^2_{RS}			−0.013	−0.07				
Reversed × $\rho_{1,2}$			0.006	0.03				
Reversed × $\rho_{1,\mathrm{RS}}$			−0.031	−0.17				
Reversed × $\rho_{2,\mathrm{RS}}$			−0.031	−0.16				
$\sigma^2_{\mathrm{RS}} \times \rho_{1,2}$			−0.055	−0.06				
$\sigma^2_{\mathrm{RS}} \times \rho_{1,\mathrm{RS}}$			0.073	0.07				
$\sigma^2_{\mathrm{RS}} \times \rho_{2,\mathrm{RS}}$			0.075	0.07				
$\rho_{1,2} \times \rho_{1,\mathrm{RS}}$			−0.075	−0.07				
$\rho_{1,2} \times \rho_{2,\mathrm{RS}}$			−0.072	−0.07				
R^2		0.15		0.91		0.00		0.89

Note. All predictor variables were centered. All *SE*s for paramters $b \leq .001$.

The second model revealed several interaction effects. The most important one was the interaction between the two attribute-ARS correlations ($\rho_{1,\mathrm{RS}} \times \rho_{2,\mathrm{RS}}$), which is also depicted in Figure 4. The correlation of interest was overestimated if the attribute-ARS relationships were either both positive or both negative, and the correlation of interest was underestimated if the attribute-ARS relationships were of opposite sign. However, bias was small if at least one of the attributes was unrelated to ARS. Apart from that, reverse-keyed items buffered against the detrimental effect of an attribute-ARS relationship. However, this holds only for one attribute at a time and not for their interaction: The three way interaction (Rev × $\rho_{1,\mathrm{ARS}} \times \rho_{2,\mathrm{ARS}}$) did not explain additional variance. Furthermore, all effects became stronger the more ARS variance was in the data as revealed by the respective interactions.

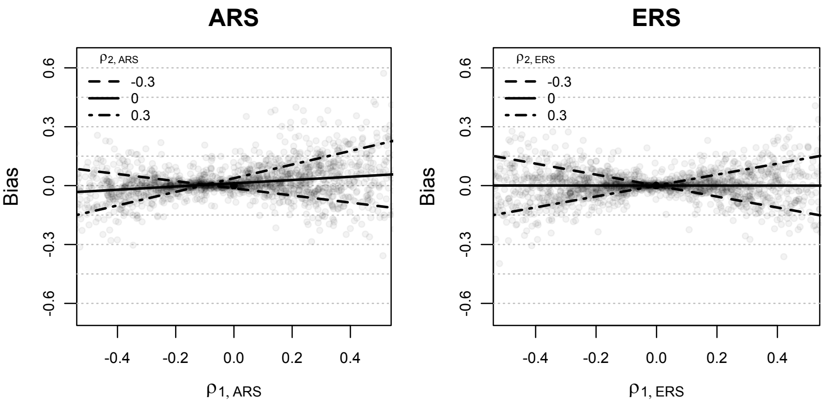


FIGURE 4: Effect of ARS and ERS, respectively, on the correlation of two scale scores. Plotting region was restricted to $|\rho_{1,\,\mathrm{RS}}| < .5$ and a subset of 2,000 replications.

Extreme Response Style

The previous simulation was repeated focusing now on ERS, and the results are displayed in Table 2. Both regressions indicated that bias was, on average, virtually zero. Moreover, none of the first-order predictors explained a substantial amount of variance. However, the interaction between the two attribute-ERS relationships was again large, which mirrors the quadratic effect (of $\rho_{1,\,\mathrm{ERS}}$) observed in the scenario before. If all three intercorrelations were positive, people high (low) on both attributes gave more (less) extreme responses, which shifted these responses upwards. If both attribute-ERS relationships were negative, however, responses were shifted downwards. In both cases, the shared variance among items was inflated leading to on overestimation of the correlation between the two attributes. Contrarily, attribute-ERS correlations that were of opposite sign resulted in inverted patterns across the two attributes (upwards shift on one attribute and downwards shift on the other), which deflated the shared variance among items. Most important, however, was the effect that bias was virtually zero if one attribute-ERS relationship was close to zero. Furthermore, average bias did not exceed values of .08 for moderate attribute-ERS relationships ($|\rho| < .3$).

Estimating Respondents' Scale Scores in the Presence of Response Styles

In the final scenario, the effect of response styles on respondents' scale scores was investigated. The previous results already revealed that response styles can affect the relationship between observed and true scores (i.e., the reliability)—and this is of course due to distorted scores of respondents. The following simulations were intended to show a more fine-grained picture at the level of individual scores, because these are often the final goal in many situations. This made it necessary, to reduce the complexity of the simulation design in order to keep the presentation of the results concise. Therefore, an extreme condition with $\sigma^2_{\text{RS}} = 1$ was contrasted with a situation were response styles were absent (i.e., $\sigma^2_{\text{RS}} = 0$). Furthermore, it was decided to focus on reverse-keyed items, because this is the variable that can directly be controlled by the researcher. The effect of zero, two, and four reverse-keyed items was investigated for ARS (and arbitrarily set to four for ERS). Apart from that, 10 items, five categories, 200 persons, $\mu_{\text{RS}} = 0$, and $\rho_{1,\text{RS}} = 0$ was specified for each sample. Given the reduced number of conditions, only 10,000 replications were run in each simulation.

The results (i.e., correlations, TPR, FPR, and accuracy) are depicted in Figure 5. Even in the absence of response styles (depicted in gray), there was some natural discrepancy between the observed scales scores, $\bar{x}_1$, and the true person parameters, θ_1, due to the unreliable measurement with only 10 five-point items ($r = .93$). This was also reflected in the non-perfect accuracy, TPR, and FPR.

The effect of ERS and ARS, respectively, is mirrored in the difference between the response style condition (displayed in black) and the baseline condition (displayed in gray). In the uppermost panel of Figure 5, the influence of ERS is depicted, and the results indicated that ERS was problematic with respect to the TPR when selecting the highest performing individuals. For example, the TPR dropped from .84 to .81 at $c = .80$ (i.e., the uppermost 20% of a sample are selected). The accuracy indicated that ERS was most influential in the mildly extreme areas of the scale. In this range, ERS may make a difference between a 4- and a 5-response (or a 1 and a 2). In the outermost areas, the attribute level is so high or low making ERS less influential. Similarly, in the center of the scale, only very extreme ERS levels have the potential to alter responses in categories 2, 3, and 4.

The results for ARS with four, two, and zero reverse-keyed items are displayed in the three lower panels of Figure 5. All three measures were impaired in the presence of ARS, the more so the less reverse-keyed items were used. For example, with zero reverse-keyed items at $c = .80$, the TPR was only .76 (compared to .84 in the baseline condition), the FPR increased to .08 (compared to .06), and the accuracy was .88 (compared to .92). However, this effect was substantially reduced when using two or even four reverse-keyed items. In the latter case, ARS had virtually no effect at all. Note that the slight asymmetry in the impact of ARS (i.e., higher impact in the upper range of the scale) is simply due to an odd number of categories (ARS contrasts two agree-categories with three non agree-categories) and would disappear with an even number of categories.

Taken together, even though only an extreme condition—response style variance equal to content variance—was investigated herein, the effect of response

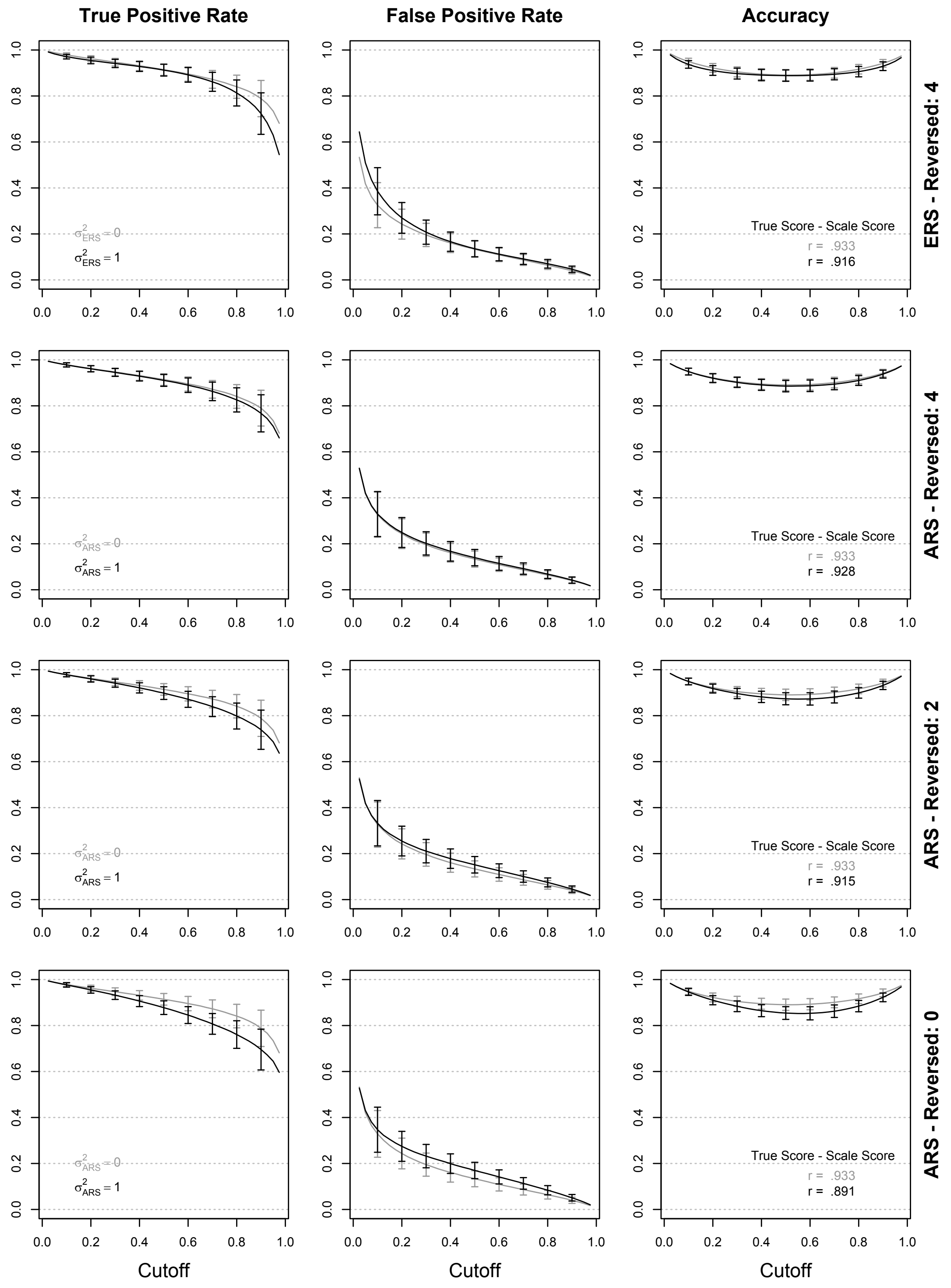


Figure 5: Influence of ERS and ARS on the classification of respondents, which is mirrored in the difference between the baseline condition without response styles (in gray) and the response style condition (in black). The lines represent the mean across all replications at a given cutoff value, error bars represent standard deviations.

styles was once again rather minor. This was especially true with respect to ERS and with respect to ARS controlled for by reverse-keyed items.

An Illustrative Example

An empirical data set from Jackson (2012) was analyzed in order to illustrate the effects of response styles in real data and to check whether the parameter values chosen in the simulations were reasonable. Respondents that were older than 80 ($n = 12$) or with unclear sex ($n = 56$) were excluded. Furthermore, 23 cases were removed because these persons showed no variability in the chosen response option across more than 25 subsequent items. The final data set included 8,745 persons who provided responses to 50 Big Five items. Openness, Conscientiousness, Extraversion, Agreeableness, and Emotional Stability were measured by ten 5-point items each (including 3, 4, 5, 4, and 8, respectively, reverse-coded items). Two models were fit to the data using a partial credit model parametrization[6]: Model 1 comprised one dimension for each of the five scales (between-item multidimensionality), and Model 2 included two additional dimensions for ARS and ERS, respectively, that where each measured by all 50 items. The model was fit using the R package **TAM** (Kiefer, Robitzsch, & Wu, 2015) employing Quasi-Monte Carlo integration with 5,000 nodes. For the sake of brevity, only model-based results (rather than raw score analyses) are reported in the following.

Model 2 had 13 parameters more than Model 1 (two variances, 11 covariances) and was clearly superior in terms of model fit (e.g., BIC-values of 1,130,022 for M1 and 1,093,083 for M2). The estimated item parameters of the two models were highly similar, $r > .99$, with a mean absolute difference (MAD) of .09. The correlations of the five pairs of corresponding person parameters (EAP) estimated by the two models were $r = .88$, $r = .95$, $r = .97$, $r = .90$, and $r = .96$; MADs ranged between 0.14 and 0.26. In Model 2, the estimated variances of the Big Five dimensions were 0.59, 0.54, 1.20, 0.65, and 0.89, and those values were on average .04 smaller compared to Model 1. Furthermore, the estimated variance of ARS was 0.14 and that of ERS was 1.02. The latent intercorrelations of the Big Five dimensions in Model 2 ranged from $-.04$ to .46; the differences between these correlations and those from Model 1 ranged from $-.03$ to .04 with an average of .02 in absolute terms. The correlations between ARS and the Big Five dimensions ranged from $-.09$ to .11, and the correlations between ERS and the Big Five dimensions ranged from $-.19$ to .04. The estimated correlation between ARS and

[6] A rating scale model parametrization fit the data worse but did not affect the interpretation of the results. Few if any differences regarding the coefficients reported herein were observed in the second or third decimal place.

ERS was $r = .26$. Finally, in Model 2, the estimated EAP reliabilities for the Big Five ranged from .77 to .89 with an average of .83. Those values were on average .02 (between .01 and .04) smaller than those from Model 1 indicating that the reliability was slightly overestimated when response styles were ignored.

In summary, these results indicated that controlling for response styles increased model fit and led to different parameter values. However, these differences were rather small. Apart from that, the response style variances and covariances were in the range of the values chosen in the simulation above (with the only, small exception being σ^2_{ERS}).

Discussion

There is an increasing interest in response styles, and many models to measure and control for response styles have been developed. The justification for this research activity is—partly—the belief that not taking response styles into account would distort self-report data, which are used all over the place in (social) science. The goal of the present research was to scrutinize this belief with a focus on applied settings and, in turn, to take a more systematic look at the role of response styles.

Therefore, a simulation study was carried out for the two most prominent response styles, ARS and ERS, and for three different scenarios: one looking at Cronbach's alpha, one looking at correlations, and one looking at respondents' scores. These scenarios were selected to resemble typical situations of applied data analysis, where response styles are often ignored—either because response styles are believed to be negligible or because methodological control cannot be realized for one reason or another.

While the generated data were analyzed with everyday methods, they were simulated from a sophisticated, but straightforward IRT model. Wetzel and Carstensen (2015) extended the idea of within-item multidimensionality in the polytomous Rasch model to dimensions that are not related to content but to response styles. The only difference to traditional within-item multidimensionality is that the response style dimension receives weights that are different from the traditional, ordinal coding. This model was well suited for the present simulations, because it is highly flexible and different response styles and/or multiple attributes can be incorporated. Moreover, the underlying notion of response styles is similar to existing approaches.

The results were twofold. On the one hand, bias was large when the attribute of interest was correlated with response style, and bias got extreme for large correlations. Such attribute-response style relationships might perhaps explain empirical findings of a notable impact of response styles. However, correlations outside ±.20 were not observed in the empirical example presented herein and may in general be the exception rather than the rule. Moreover, if the correlation really is in the range of .40, .60, or even higher, the question arises what the items at hand actually measure and whether the problem may be socially desirable responding rather than ARS (cf. Paunonen & LeBel, 2012). In such situations, self-reports might not be a sensible way of data collection.

On the other hand, bias was small or even negligible in a large range of conditions. This holds especially if the attribute-response style relationship was small. Moreover, bias was lower when more reverse-keyed items were used (for ARS), when the attribute-attribute correlation was higher, and when response style variance was smaller. For example, in the conditions in Figure 2 where response styles were unrelated to the attribute(s), bias hardly exceeded levels of .05 or even .02 if at least two reverse-coded items were used. In summary, the findings are in line with previous work finding only small effects as long as the attribute-response style relationship is small (Ferrando & Lorenzo-Seva, 2010; Johnson & Bolt, 2010; Savalei & Falk, 2014; Wetzel et al., 2016).

The presented empirical example supported the interpretation that response styles can introduce bias, but that this bias is rather small and unlikely to alter results completely. Moreover, the example showed that the parameter values chosen for the simulation were reasonable and definitely not understated.

The issue of an attribute-response style relationship brings up further questions about causation and the nature of response styles themselves. Let's look at two examples with ARS. First, if a bivariate relationship between an attribute and ARS is observed, this may be simply due to a common cause or confounder (e.g., cultural background) whilst the bivariate relationship is in fact non-existing. Thus, a correct model would include the confounder but not necessarily ARS. Second, two independent attributes may both causally influence ARS—then called a collider. If ARS is wrongly included in the model, a spurious relationship between the two attributes may result. These examples highlight that a much deeper understanding of response styles and their causes and causal effects is needed in order to evaluate the impact of attribute-response style relationships.

If a rule of thumb should be derived from the present results, ⅓ of reverse-keyed items are probably a good way to control ARS. There was no evidence that a fully balanced scale would further improve the results markedly, at least if the attribute-response style correlation was reasonably small. However, it should be noted that the use of reverse-keyed items may have downsides, and there is ample literature on that topic (cf. Weijters, Baumgartner, & Schillewaert, 2013). Apart from that, the number of reverse-keyed items had no effect on the bias caused by

ERS. This is not surprising given that the definition of ERS is independent of the direction of an item.

As with every simulation study, the generalizability of the results depends on the external validity (a) of the simulation model, (b) of the parameter values (fixed or varied), and (c) of the analysis model. First, the chosen model allowed for a very natural implementation of ERS—a tendency to select the endpoints—and ARS—a tendency to agree. Moreover, the model is highly similar to existing approaches and there is no reason to assume that a different simulation model (e.g., Johnson & Bolt, 2010) would lead to fundamentally different results. Furthermore, differences between the chosen rating scale approach and a partial credit approach would probably cancel each other out across items and replications. Second, the chosen parameter values seemed plausible given the empirical example. Moreover, the regression results make it straightforward to plug in values (e.g., $\sigma^2_{\text{RS}} = 2$) that were not covered herein. And, a wide range of conditions was realized by randomly sampling from the independent variables instead of restricting the study to, say, three levels of every factor. This, in turn, allowed to uncover quadratic and interaction effects. Third, the analyses focused on bias, which was based on partialing response style from the measure of interest. If the attribute and response style were correlated, this led to the fact that also attribute variance was—wrongly—partialed out inflating the amount of bias, the more so the stronger the correlation was. Thus, the extreme levels of bias (e.g., for $\rho = .5$) are probably a (too) pessimistic estimate. Apart from that, the analyses focused on only three scenarios, but the results translate to more complex situations, for example, when more than two attributes are investigated in a structural model.

Different outcomes, such as factor structure, model fit, threshold and loading parameters, or higher-order moments were not covered herein and remain a route for further research. Moreover, the relationship among different response styles and the effect of multiple response styles at a time may be of interest in future studies. Apart from that, this study focused on the effect of ignoring response styles in raw score-analyses; whether and how response styles can be controlled using appropriate (model-based) approaches is a different question (see, e.g., Wetzel et al., 2016).

In summary, the present results suggest that the impact of response styles in applied settings is probably better described by a molehill than a mountain. The analyses demonstrated the importance of reverse-keyed items to control for the negative influence of ARS. The future will show whether the gap between the applied camp and the methods camp can be bridged such that practitioners take response styles into account where necessary and that psychometricians develop and refine the tools required to do so.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23. doi:10.1177/0146621697211001

Aichholzer, J. (2013). Intra-individual variation of extreme response style in mixed-mode panel studies. *Social Science Research*, *42*, 957–970. doi:10.1016/j.ssresearch.2013.01.002

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814

Bentler, P. M. (2016). Covariate-free and covariate-dependent reliability. *Psychometrika*, *81*, 907–920. doi:10.1007/s11336-016-9524-y

Billiet, J. B. & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628. doi:10.1207/S15328007SEM0704_5

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

Bolt, D. M. & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*, 814–833. doi:10.1177/0013164410388411

De Boeck, P. & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. doi:10.18637/jss.v048.c01

Eid, M. & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*, 20–30. doi:10.1027//1015-5759.16.1.20

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

Ferrando, P. J. & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 427–448. doi:10.1348/000711009X470740

Finn, J. A., Ben-Porath, Y. S., & Tellegen, A. (2015). Dichotomous versus polytomous response options in psychopathology assessment: Method or meaningful variance? *Psychological Assessment*, *27*, 184–193. doi:10.1037/pas0000044

Hankin, R. K. S. (2005). Recreational mathematics with R: Introducing the "magic" package. *R News*, *5*(1), 48–51.

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125. doi:10.1177/014662169602000201

Heide, M. & Grønhaug, K. (1992). The impact of response styles in surveys: A simulation study. *Journal of the Market Research Society*, *34*, 215–230.

Jackson, A. (2012). IPIP Big Five personality test answers [Data file]. doi:10.6084/m9.figshare.96542

Jin, K.-Y. & Wang, W.-C. (2014). Generalized IRT models for extreme response style. *Educational and Psychological Measurement*, *74*, 116–138. doi:10.1177/0013164413498876

Johnson, T. R. & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*, 92–114. doi:10.3102/1076998609340529

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*, 161–177. doi:10.1080/00273171.2013.866536

Kiefer, T., Robitzsch, A., & Wu, M. (2015). TAM: Test analysis modules (Version 1.11-0). Retrieved from https://CRAN.R-project.org/package=TAM

Meiser, T. & Machunsky, M. (2008). The personal structure of personal need for structure: A mixture-distribution Rasch analysis. *European Journal of Psychological Assessment*, *24*, 27–34. doi:10.1027/1015-5759.24.1.27

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Paunonen, S. V. & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology*, *103*, 158–175. doi:10.1037/a0028165

Plieninger, H. & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899. doi:10.1177/0013164413514998

R Core Team. (2014). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability: Vol. 4. Contributions to Biology and Problems of Medicine* (pp. 321–333). Berkeley, CA: University of California. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512895

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin*, *63*, 129–156. doi:10.1037/h0021888

Savalei, V. & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*, 407–424. doi:10.1080/00273171.2014.931800

Schimmack, U., Böckenholt, U., & Reisenzein, R. (2002). Response styles in affect ratings: Making a mountain out of a molehill. *Journal of Personality Assessment*, *78*, 461–483. doi:10.1207/S15327752JPA7803_06

Trautmann, H., Steuer, D., Mersmann, O., & Bornkamp, B. (2014). truncnorm: Truncated normal distribution (Version 1.0-7). Retrieved from http://CRAN.R-project.org/package=truncnorm

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*, 195–217. doi:10.1093/ijpor/eds021

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Statistics and Computing. DOI: 10.1007/978-0-387-21706-2. New York, NY: Springer.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*, 320–334. doi:10.1037/a0032121

Weijters, B., Cabooter, E. F. K., & 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*, 236–247. doi:10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*, 105–121. doi:10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*, 96–110. doi:10.1037/a0018721

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*, 304–324. doi:10.1177/0013164415591848

Wetzel, E. & Carstensen, C. H. (2015). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment*. Advance online publication. doi:10.1027/1015-5759/a000291

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178–189. doi:10.1016/j.jrp.2012.10.010

Appendix

Varying the Number of Categories and the Impact of the Weights

Researchers are regularly confronted with the question about the optimal number of categories for their instrument, and there is ample research trying to answer this question from different perspectives (e.g., Finn, Ben-Porath, & Tellegen, 2015). At first sight, it seems plausible to add to this literature with the present simulation study. Therefore, the number of categories was varied between 3 and 7 in a simulation focusing on the effect of ARS on Cronbach's alpha with a minimal setup ($\mu_{\text{ARS}} = 0; \sigma^2_{\text{ARS}} = 1; \rho_{1,\,\text{ARS}} = 0$). At first sight, more categories led to less bias ($b = -.01$). However, there is a confounding effect between the number of categories and the weights in the **B** matrix: With increasing categories, the content-related weights increase in size (e.g., from $[0, 1, 2]'$ to $[0, 1, 2, 3]'$), whereas the ARS weights are always fixed to zeros and ones (e.g., $[0, 0, 1]'$ and $[0, 0, 1, 1]'$). Thus, more categories did not lead to less bias for substantive reasons, but simply for the reason that the relative size of the ARS weights (i.e., the impact of ARS) decreased. For illustration, the ARS weights were set to $^{K}\!/_{2}$ in a follow-up simulation (e.g., $[0, 0, 1]'$ and $[0, 0, 1.5, 1.5]'$). Then, the effect of the number of categories on bias changed sign ($b = .01$). Thus, the confounding effect between the number of categories and the response style weights makes it impossible to draw conclusions about the relationship between the number of categories and response styles.

This confounding effect would play a role in all simulations reported herein. Moreover, it is present whenever weights for response styles are used and different numbers of categories are compared.

Apart from that, the described mechanism also applies to the comparison of different scoring schemes for a fixed number of categories. For example, weights of $(2, 1, 0, 1, 2)'$ for ERS instead of $(1, 0, 0, 0, 1)'$ may be seen just as valid. However, increasing the weights would also artificially increase the impact of ERS making the results incomparable. This is also mirrored in the fact that, in the empirical illustration, σ^2_{ERS} dropped from a value of 1.02 to 0.35 if the ERS weights were changed to $(2, 1, 0, 1, 2)'$.

A New Model for Acquiescence at the Interface of Psychometrics and Cognitive Psychology

Hansjörg Plieninger and Daniel W. Heck
University of Mannheim

Abstract

When measuring psychological traits, one has to consider that respondents often show content-unrelated response behavior in answering questionnaires. To disentangle the target trait and two such response styles, extreme responding and midpoint responding, Böckenholt (2012) developed an item response model based on a latent processing tree structure. We propose a theoretically motivated extension of this model to also measure acquiescence, the tendency to agree with both regular and reverse-coded items. Substantively, our approach builds on multinomial processing tree (MPT) models that are used in cognitive psychology to disentangle qualitatively distinct processes. Accordingly, the new model assumes a mixture distribution of affirmative responses, which are either determined by the underlying target trait or by acquiescence. In order to estimate the model parameters, we rely on Bayesian hierarchical estimation of MPT models. In simulations, we show that the model provides unbiased estimates of response styles and the target trait, and we compare the new model and Böckenholt's model in a model recovery study. An empirical example from personality psychology is used for illustrative purposes.

We would like to thank Morten Moshagen, Benjamin Hilbig, and Ingo Zettler for generously providing their data for secondary analysis as well as Thorsten Meiser for helpful comments on a previous version of this manuscript.

Introduction

Questionnaires with Likert-type response formats are widely used to assess various constructs such as personality variables, mental disorders, or attitudes towards products, teachers, or co-workers. Despite their widespread application, however, severe concerns have been raised about the validity of Likert-type data because of response styles. Such response styles are defined as systematic preferences of respondents for specific response categories that cannot be explained by the item content. Three prominent response styles are the "tendency to use positive response categories (acquiescence response style, or ARS), ... the midpoint response category (midpoint response style, or MRS), and extreme response categories (extreme response style, or ERS)" (Weijters, Geuens, & Schillewaert, 2010b, p. 96). Previous research showed that response styles are stable across time and domains and can be measured as trait-like constructs (e.g., Danner, Aichholzer, & Rammstedt, 2015; Weijters, Geuens, & Schillewaert, 2010a, 2010b; Wetzel, Carstensen, & Böhnke, 2013).

Early psychometric models for response styles usually focused on a single response style, for example, mixture distribution Rasch models for ERS (e.g., Rost, Carstensen, & von Davier, 1997) or factor models for ARS (e.g., Billiet & McClendon, 2000). In recent years, models that account for more than one response style have been proposed (e.g., Johnson & Bolt, 2010); for example, Böckenholt (2012) proposed so-called item-response-tree (IR-tree) models to account for both MRS and ERS. This model has the appeal that it assumes a psychologically meaningful tree-structure of the underlying processes. In the present manuscript, we extend this model to acquiescence by building on multinomial processing tree (MPT) models from cognitive psychology and recent computational advances in Bayesian hierarchical estimation of these models. In the remainder of the Introduction, we develop the proposed model on the foundation of MPT models, hierarchical extensions of MPT models, and Böckenholt's model and discuss the novel theoretical account of acquiescence as a mixture distribution. Next, two simulation studies address parameter and model recovery, respectively. Finally, an empirical example is used to illustrate the benefits of the proposed approach.

Multinomial Processing Tree Models

Multinomial processing tree models are widely used in cognitive and social psychology to disentangle a finite number of qualitatively distinct processes that are assumed to result in identical responses (Erdfelder et al., 2009; Hütter & Klauer, 2016; Riefer & Batchelder, 1988). For example, in a typical recognition-memory paradigm, respondents first learn a list of words and then have to categorize words either as *OLD* or *NEW* in the subsequent recognition test. Figure 1 shows a specific MPT model for this paradigm, the so-called *one-high-threshold model* (Green & Swets, 1966), which assumes that responses emerge from two qualitatively distinct latent processes: When presented with a learned/old item, respondents either enter a state of recollection certainty with probability r (and respond *OLD* accordingly), or they enter a state of recollection uncertainty with probability $1 - r$. In the latter case, no information about the test item is available and hence participants have to guess *OLD* or *NEW* with probabilities g and $1 - g$, respectively. When presented with a new item, it is assumed that respondents are always in an uncertainty state and have to guess *OLD* or *NEW*, again with probabilities g and $1 - g$, respectively.

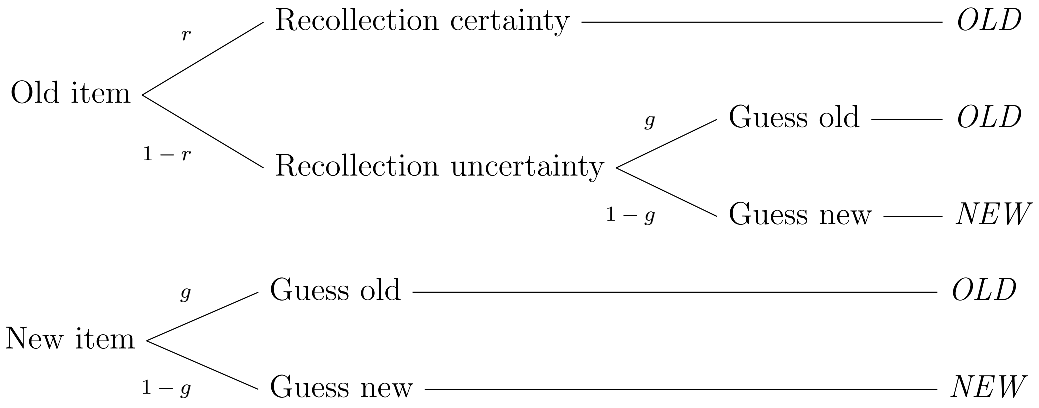


FIGURE 1: Example of a multinomial processing tree model, namely, the one-high-threshold model of recognition memory.

Statistically, the model entails two free parameters ξ_p $(p = 1, \ldots, P)$. The parameter $\xi_1 = r$ measures memory strength (i.e., the probability that a learned item crosses the recognition threshold), whereas the parameter $\xi_2 = g$ measures response bias (i.e., the probability to guess *OLD* in the absence of any memory signal). Based on these parameters, the model equations for the observable response categories are given as:

$$Pr(\text{OLD} \mid \text{old item}) = r + (1 - r)g \tag{1}$$

$$Pr(\text{NEW} \mid \text{old item}) = (1 - r)(1 - g) \tag{2}$$

$$Pr(\text{OLD} \mid \text{new item}) = g \tag{3}$$

$$Pr(\text{NEW} \mid \text{new item}) = 1 - g. \tag{4}$$

As illustrated in Figure 1, the expected probability for each processing branch m $(m = 1, \ldots, M)$ is given by multiplying the corresponding parameters; for example, the probability of guessing *OLD* for old items is simply $(1-r)g$. Since branches are assumed to be disjoint, branch probabilities resulting in identical responses are summed to obtain expected category probabilities; for example, an *OLD*-response to an old item may either result from recollection or from guessing *OLD* and is therefore modeled as $r + (1-r)g$. Note that, statistically, MPT models assume a mixture of different processes where the mixing probabilities are constrained substantively by a tree structure as that in Figure 1.

Given the expected category probabilities, the response frequencies $\boldsymbol{x}$ across the k observable categories ($k = 1, \ldots, K$) are assumed to follow a product-multinomial distribution. Hence, the likelihood of an MPT model can be expressed in general form as[1]

$$Pr(\boldsymbol{x} \mid \boldsymbol{\xi}) \propto \prod_{k=1}^{K} \left(\sum_{m=1}^{M} c_m \prod_{p=1}^{P} \xi_p^{v_{mp}} (1-\xi_p)^{w_{mp}} \right)^{x_k}, \tag{5}$$

where v_{mp} and w_{mp} count how often the parameters ξ_p and $(1-\xi_p)$ occur in the branch m, respectively, and c_m represents possible constants (e.g., due to constraints such as $g = .50$; Hu & Batchelder, 1994). Based on a vector of observed response frequencies $\boldsymbol{x} = (x_1, \ldots, x_K)$, parameters can easily be estimated by maximizing the likelihood in Equation 5 using an EM algorithm (Hu & Batchelder, 1994). In cognitive psychology, responses are often assumed to be independently and identically distributed across both persons and items, which allows for aggregation to obtain a sufficient number of observations for parameter estimation. Note that the one-high-threshold model uses two parameters to model two non-redundant model equations, which results in a saturated model (i.e., $df = 0$).

Hierarchical MPT Models

As noted above, MPT modeling in cognitive psychology has often rested on the assumption of homogeneity across both persons and items. However, this restrictive assumption has been questioned in recent years, and this was accompanied by the call for models that take heterogeneity of persons and/or items into account (e.g., Rouder & Lu, 2005). Recently, Klauer (2010) and Matzke, Dolan, Batchelder, and Wagenmakers (2015) have developed hierarchical MPT models that allow for person- and/or item-specific effects, thereby overcoming the need to aggregate the data (see also Smith & Batchelder, 2010).

In hierarchical MPT extensions (Klauer, 2010; Matzke et al., 2015), the parameters ξ_{pij} are allowed to vary over both persons (indexed $i = 1, \ldots, I$) and

[1] We only provide the proportional likelihood function without the product of multinomial constants to enhance readability and to reduce the number of subscripts.

items (indexed $j = 1, \ldots, J$). For each person-item combination, the parameters $(\xi_{1ij}, \ldots, \xi_{Pij})$ define an MPT likelihood function identically as in Equation 5. In addition, however, the parameters ξ_{pij} are reparameterized using an IRT-like structure with additive person- and item-effects. Since the MPT parameters are defined as probabilities on the range $[0, 1]$, the parameters are first mapped to the real line using the probit-link function $\Phi^{-1}(\xi_{pij})$ (i.e., the inverse cumulative distribution function of the standard normal distribution). Then, on the probit scale, the person ability parameter θ_{pi} and the item difficulty parameter β_{pj} are assumed to combine additively,

$$\xi_{pij} = \Phi\left(\theta_{pi} - \beta_{pj}\right). \tag{6}$$

For instance, when modeling the memory parameter r in the one-high-threshold model in such a way, r_{ij} is assumed to increase for participants with better memory (high θ_{pi}) and with easy-to-remember words (low β_{pj}).

Put differently, each MPT parameter ξ_p is first modeled as the dependent variable of a binary IRT model (i.e., a probit-link IRT or Rasch model). Then, the MPT parameters in Equation 6 can be plugged separately into the MPT likelihood in Equation 5 resulting in a hierarchical MPT model. Overall, this combination of psychometric measurement models with cognitive process models provides a powerful framework that has received considerable attention in cognitive psychology but not yet in psychometrics. With the present work, we aim at (a) raising the awareness for such modeling approaches previously also termed *cognitive psychometrics* (Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002), and (b) developing a novel, cognitively-inspired model for acquiescence, as derived in the following sections.

The IR-Tree Model for Response Styles

Böckenholt (2012), as well as De Boeck and Partchev (2012), developed the class of IR-tree models. These models can be subsumed under the framework of hierarchical MPT models, even though they were developed independently. Herein, we will focus on one exemplar of IR-tree models, namely, a response style model for questionnaire items using an ordinal, symmetric 5-point response format (henceforth called the *Böckenholt Model*). Usually, a set of such items is analyzed using a one-factor model like the partial credit model (Andrich, 1978). However, in the IR-tree approach, it is assumed that three qualitatively distinct processes account for the observed responses (see Figure 2): First, persons may enter an MRS stage with probability m_{ij} (i.e., $\xi_1 = m$) and give a midpoint response. The complementary stage is entered with probability $1-m_{ij}$; then, a latent state of a high level of the target trait is entered with probability t_{ij} eventually leading to agreement or a low level is entered with probability $1-t_{ij}$ eventually leading to disagreement. For example, if the items are designed to measure happiness, the states may be interpreted in terms of happy versus unhappy. Finally, an ERS stage is entered with probability e_{ij} leading to a *strongly agree*-response in case of agreement and a *strongly disagree*-response in case of disagreement. The complementary stage is entered with probability of $1-e_{ij}$ leading to moderate *agree*- and *disagree*-responses, respectively.

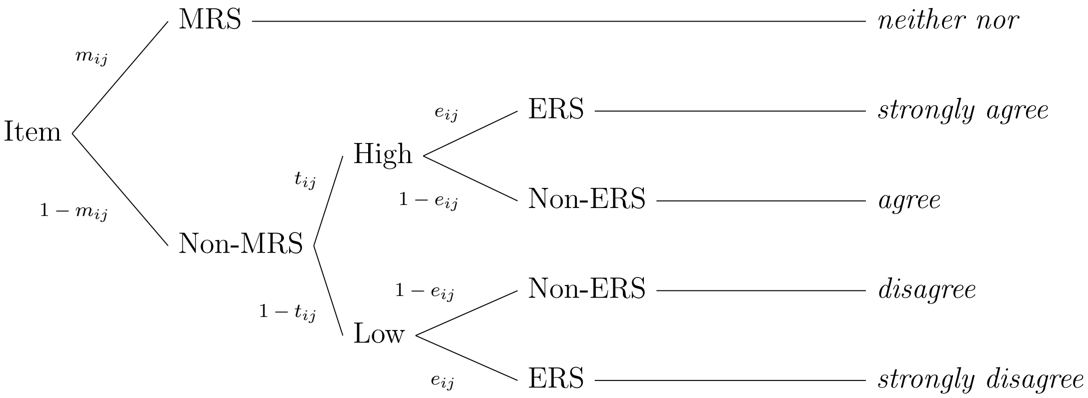


FIGURE 2: Böckenholt's (2012) IR-tree model for a 5-point item accounts for midpoint response style (MRS) and extreme response style (ERS) besides the target trait (High/Low).

The three MPT parameters m, e, and t are reparameterized as follows[2]:

$$m_{ij} = \Phi(\theta_{mi} - \beta_{mj}) \tag{7}$$
$$e_{ij} = \Phi(\theta_{ei} - \beta_{ej}) \tag{8}$$
$$t_{ij} = \Phi(\theta_{ti} - \beta_{tj}). \tag{9}$$

Substantively, each person is assumed to have three latent traits $\boldsymbol{\theta}_i = (\theta_{mi}, \theta_{ei}, \theta_{ti})'$ and each item is modeled using three difficulty parameters $\boldsymbol{\beta}_j = (\beta_{mj}, \beta_{ej}, \beta_{tj})'$. Individual differences with respect to the target trait (e.g., happiness) are measured by θ_{ti}. Moreover, individual differences in response styles—which are both consistently found in the literature (e.g., Johnson & Bolt, 2010) and sometimes of theoretical interest themselves (e.g., Zettler, Lang, Hülsheger, & Hilbig, 2016)—are measured by θ_{mi} and θ_{ei}. Besides these person parameters, the item difficulty β_{tj} measures how likely it is to agree with an item. However, items may also vary in their response-style-related difficulty (e.g., De Jong, Steenkamp, Fox, & Baum-

[2] For the sake of readability, the IRT parameters θ and β are indexed using not the cardinal number but rather the symbol of the respective parameter ξ_p (e.g., θ_{mi} is used instead of θ_{1i}).

gartner, 2008). For example, some items may elicit few extreme responses (i.e., high β_{ej}), or some items may elicit many midpoint responses (i.e., low β_{mj}).

According to the tree in Figure 2, and similar as in MPT models, the probability of a response is given by multiplying all the probabilities along the corresponding branches. For instance, the probability to strongly agree with an item is given by

$$Pr(x_{ij} = 5 \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}_j) = (1 - m_{ij}) t_{ij} e_{ij}. \tag{10}$$

Böckenholt (2012), as well as De Boeck and Partchev (2012), proposed to estimate the model using existing maximum likelihood software for standard multidimensional IRT models. For this purpose, an observed response is recoded into three binary pseudoitems that correspond to the outcomes of the three latent stages. The first pseudoitem encodes whether the middle category was chosen or not, the second whether the respondent agreed or disagreed, and the third whether an extreme or a moderate response was given (midpoint responses are coded as missing by design on the last two pseudoitems). Based on this recoding, the Böckenholt Model can be fit similarly as a standard, three-dimensional binary IRT model. It is important to note that this method is limited to IR-tree models in which each response category is reached by a single processing path (Böckenholt, 2012). Statistically, this limitation is due to the mixture structure for the expected response probabilities that is implied if multiple branches lead to the same response category (see MPT models above).

In summary, instead of assuming a single latent dimension or process, the IR-tree model assumes three qualitatively distinct processes to account for MRS, ERS, and the target trait. Evidence for the validity of this approach comes from the theoretical derivation of the model based on underlying cognitive processes (Böckenholt, 2012) as well as from empirical data: For example, the construct validity of the three processes was demonstrated by Plieninger and Meiser (2014) in a study using extraneous style- and content-related criteria. Khorramdel and von Davier (2014) extended the model to questionnaires with multiple domains (Big Five) and were able to show that MRS and ERS are stable across different scales. For similar approaches to model response styles, see also Jeon and De Boeck (2016) or Thissen-Roe and Thissen (2013).

A Hierarchical MPT Model for Acquiescence

The Böckenholt Model is limited to two response styles, namely, ERS and MRS. We propose an extension of the model that takes ARS into account and that can be implemented as a hierarchical MPT model. The proposed *Acquiescence Model* builds on the basis of the Böckenholt Model and adds an additional processing stage to it. As shown in Figure 3, respondents are presented with an item and may first enter a "non-acquiescent stage" with probability $1 - a_{ij}$, which in turn leads to the branch of the Böckenholt Model (see Figure 2). However, with probability a_{ij}, respondents enter an "acquiescent stage" that always results in affirmative responses irrespective of the coding direction of the item and irrespective of the lower part of the tree. In other words, the Acquiescence Model assumes two distinct processes that lead to agreement with the items—respondents either agree because of the item's content (target trait) or merely due to a general tendency to provide affirmative responses (ARS). Analogously to the other MPT parameters, the ARS parameter is decomposed as follows:

$$a_{ij} = \Phi(\theta_{ai} - \beta_{aj}). \tag{11}$$

Respondents may differ in their ARS-level, which is captured by θ_{ai}, and items may elicit ARS-responses to different degrees, which is captured by β_{aj}.

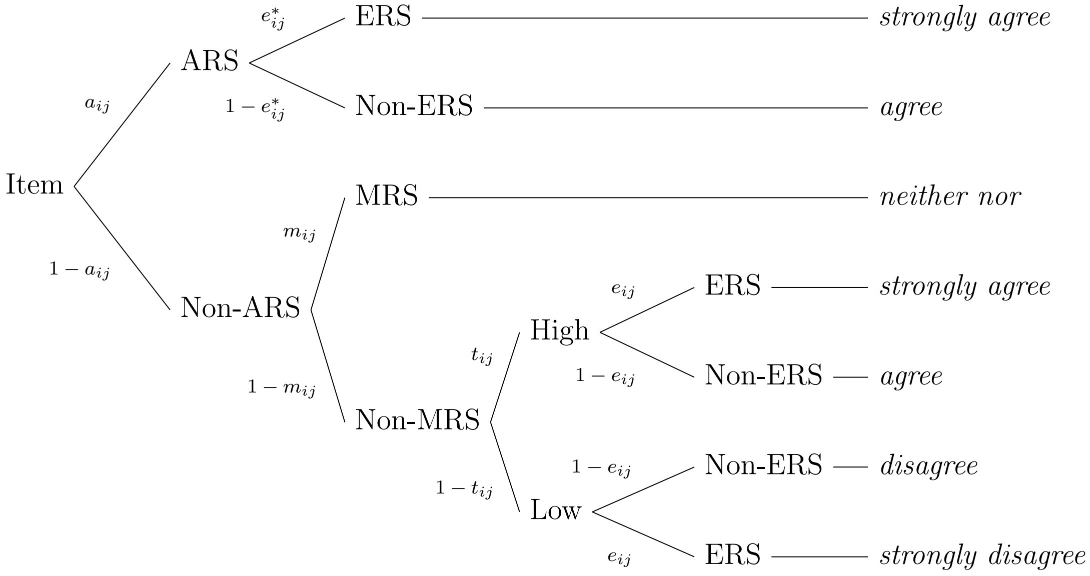


FIGURE 3: The Acquiescence Model for a regular 5-point item accounts for midpoint response style (MRS), acquiescence response style (ARS), and extreme response style (ERS) besides the target trait (High/Low). Note that multiple branches lead to agreement thereby indicating a mixture of the target trait and the ARS distribution.

Five-point items have two affirmative categories, namely, a moderate (i.e., *agree*) and an extreme one (i.e., *strongly agree*). Therefore, a further MPT parameter e^*_{ij} is necessary to model the probability of extreme responses conditional on acquiescence. The most flexible model entails a re-parameterization of this parameter as above, namely, $e^*_{ij} = \Phi(\theta_{e^*i} - \beta_{e^*j})$. However, based on theoretical considerations, we put stronger constraints on e^*_{ij} and thereby reduce model complexity. First, we set θ_{e^*i} equal to the ERS person parameter θ_{ei} from Equation 8. Substantively, we thereby assume that respondents high on ERS do not only prefer extreme over moderate categories when giving a content-related response (lower part in Figure 3), but do so similarly in case of ARS-responding (upper part). Second, we constrain all item parameters to be equal, namely, $\beta_{e^*j} = \beta_{e^*}$, which finally gives $e^*_{ij} = \Phi(\theta_{ei} - \beta_{e^*})$.

The complete model (for regular/non-reverse-coded items) is then defined by the following set of equations:

$$Pr(x_{ij} = 5 \mid \text{regular}) = a_{ij}e^*_{ij} + (1 - a_{ij})(1 - m_{ij})t_{ij}e_{ij} \tag{12}$$

$$Pr(x_{ij} = 4 \mid \text{regular}) = a_{ij}(1 - e^*_{ij}) + (1 - a_{ij})(1 - m_{ij})t_{ij}(1 - e_{ij}) \tag{13}$$

$$Pr(x_{ij} = 3 \mid \text{regular}) = (1 - a_{ij})m_{ij} \tag{14}$$

$$Pr(x_{ij} = 2 \mid \text{regular}) = (1 - a_{ij})(1 - m_{ij})(1 - t_{ij})(1 - e_{ij}) \tag{15}$$

$$Pr(x_{ij} = 1 \mid \text{regular}) = (1 - a_{ij})(1 - m_{ij})(1 - t_{ij})e_{ij}. \tag{16}$$

Importantly, any ARS model requires items of both regular (i.e., agreement is indicative of a high target-trait level) and reversed coding direction (i.e., agreement is indicative of a low target-trait level) in order to disentangle ARS and the target trait. This requirement applies to the Acquiescence Model as well. Essentially, the model assumes two distinct processing trees: The first one is depicted in Figure 3 and holds for regular items; the second one holds for reverse-coded items and is not depicted due to space considerations. It is identical to Figure 3 with a single exception in the lower branch: For reverse-coded items, the high target-trait stage eventually leads to disagreement, and the low target-trait stage eventually leads to agreement.[3]

Before estimating any MPT or IRT model, it is in general necessary to ensure that the model parameters are identifiable. Thus, it is necessary to show that different parameter values $\boldsymbol{\xi} \neq \boldsymbol{\xi}'$ imply different expected category probabilities $Pr(\boldsymbol{\xi}) \neq Pr(\boldsymbol{\xi}')$ to allow for unique parameter estimates. There are $5 - 1$ non-redundant categories in both the tree for regular and the tree for reverse-coded items, making up a total of eight. Moreover, the model is comprised of four MPT parameters (namely, m, e, a, and t), and thus a necessary condition for the identification of MPT models is fulfilled (i.e., number of free parameters $\leq$ number of non-redundant categories). In addition, local identifiability of MPT models can be assessed by ensuring that the Jacobian matrix (i.e., the matrix of the first partial

[3] Likewise, the complete model is expressed by two sets of equations. Equations 12 to 16 hold for regular items, and five additional equations are needed for reverse-coded items. These equations mirror Equations 12 to 16 with the exceptions that $(1 - t_{ij})$ is replaced by t_{ij} and t_{ij} is replaced by $(1 - t_{ij})$.

derivatives of the likelihood) has full rank for MPT parameters in the interior of the parameter space $(0,1)^P$ (Schmittmann, Dolan, Raijmakers, & Batchelder, 2010), which is the case for the Acquiescence Model (if both regular and reverse-coded items are used). Apart from the identifiability of the MPT structure of the model, the IRT parameters (namely, θ_{pi} and β_{pj}), which are used to reparameterize the MPT parameters, need to be identifiable. This is accomplished by centering the hyperpriors for $\boldsymbol{\theta}_i$ at $\mathbf{0}$ (see below; Fox, 2010, p. 86).

The presented model has some notable special cases as well as straightforward extensions. First, the Acquiescence Model reduces to the Böckenholt Model if $a = 0$ (i.e., if $(\theta_a - \beta_a) \to -\infty$). Substantively, this is the case if respondents are very low on acquiescence (low θ_{ai}) and/or if a questionnaire is very unlikely to elicit ARS-responses (high β_{aj}). Furthermore, the model reduces to a Rasch model with a probit link function if the number of categories is two and if the number of parameters $P = 1$.

Second, the model can be extended to accommodate items from more than one content domain requiring a model with more than one target trait t_d ($t_d = t_1, \ldots, t_D$). Then, each scale is modeled using separate trees, equations, and θ_d-parameters, which all increase in number by the factor D. The response style parameters should be set equal across scales mirroring the assumption that response styles are stable across content domains (Danner et al., 2015; Khorramdel & von Davier, 2014). That is, a model may contain multiple parameters θ_{t_d} (e.g., θ_{t_1} to θ_{t_6} in the empirical example below), but the three response style parameters θ_m, θ_e, and θ_a should be set equal across the D dimensions. Apart from adding multidimensional target traits, the additive IRT model for person and item effects in Equation 6 may take on more complex forms, for example, by including an item-discrimination parameter (e.g., Khorramdel & von Davier, 2014).

Theoretical Motivation

The Acquiescence Model may be seen purely as a measurement model. Substantively, however, it is in line with the theoretical account for acquiescence proposed by Knowles and Condon (1999) on the basis of the work of Gilbert (1991). Gilbert reviewed evidence on how mental beliefs are formed and compared the Cartesian and the Spinozan procedure. Descartes believed that a neutral, passive comprehension stage is followed by an active assessment stage, which leads to acceptance or rejection of an idea. Spinoza, however, believed that comprehension requires (temporal) acceptance; in a subsequent, second stage, the idea is assessed and either its acceptance is confirmed or it is “unaccepted” (Gilbert, 1991). The Cartesian procedure predicts that premature output after the first stage results in unassessed, neutral ideas, whereas the Spinozan procedure, because of its asymmetry, predicts a bias towards acceptance. Gilbert (and also Mandelbaum, 2014) made a convincing case for the Spinozan procedure and argued that it is better supported by evidence than the Cartesian procedure (however intuitive the latter may seem).

Knowles and Condon (1999) hypothesized, on the basis of the Spinozan procedure, that acquiescence is the result of a failure to fully assess an initially accepted item. This was supported by the fact that affirmative responses from persons high on ARS were faster than all other responses, and by the fact that higher ARS levels were observed under cognitive load. The Acquiescence Model builds on the Spinozan procedure: Initially, respondents agree with each item. If, unexpectedly, the response process is truncated, this results in an acquiescent response (upper part of Figure 3). However, if respondents successfully complete the assessment stage of the Spinozan procedure, they either accept (t_{ij}) or reject ($1 - t_{ij}$) an item on the basis of their target-trait level θ_{ti} and the item difficulty β_{tj}.

Mixture Versus Shift Models for Acquiescence

One of the most prominent alternative approaches to model ARS was developed within the framework of confirmatory factor analysis. A two factor model (random intercept model) is specified with (a) a target-trait factor θ^*_{ti} with item loadings λ_j that are positive for regular items and negative for reverse-coded items and (b) an ARS factor θ^*_{ai} with item loadings fixed to 1 (e.g., Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006). The model equation[4], adapted to our notation, is

$$f(x_{ij}) = \lambda_j \theta^*_{ti} + \theta^*_{ai} - \beta_j. \tag{17}$$

Note that starred versions of θ are used to distinguish the person parameters from those used in the Acquiescence Model above. The two factors in Equation 17 operate additively on the latent scale. Hence, this makes the model a *shift model* that assumes that the tendency to agree or disagree with an item follows an up- or downwards shift determined by the ARS factor (see Figure 4 for an example). Note further that different variants of this model exist, for example, some are grounded in exploratory, some in confirmatory factor analysis, and some in IRT; some allow for correlations of the two dimension and others not; some use continuous and others use discrete θ-parameters (e.g., Billiet & McClendon, 2000;

[4] Please refer to the cited references for details. Even though irrelevant to the discussion herein, note that (a) often a linear model without a link function and without category-specific item parameters is used, and that (b) inequality constraints to ensure that the λ-parameters are positive for regular and negative for reverse-coded items are often not explicitly imposed in empirical work, because the model usually converges to such a solution without them.

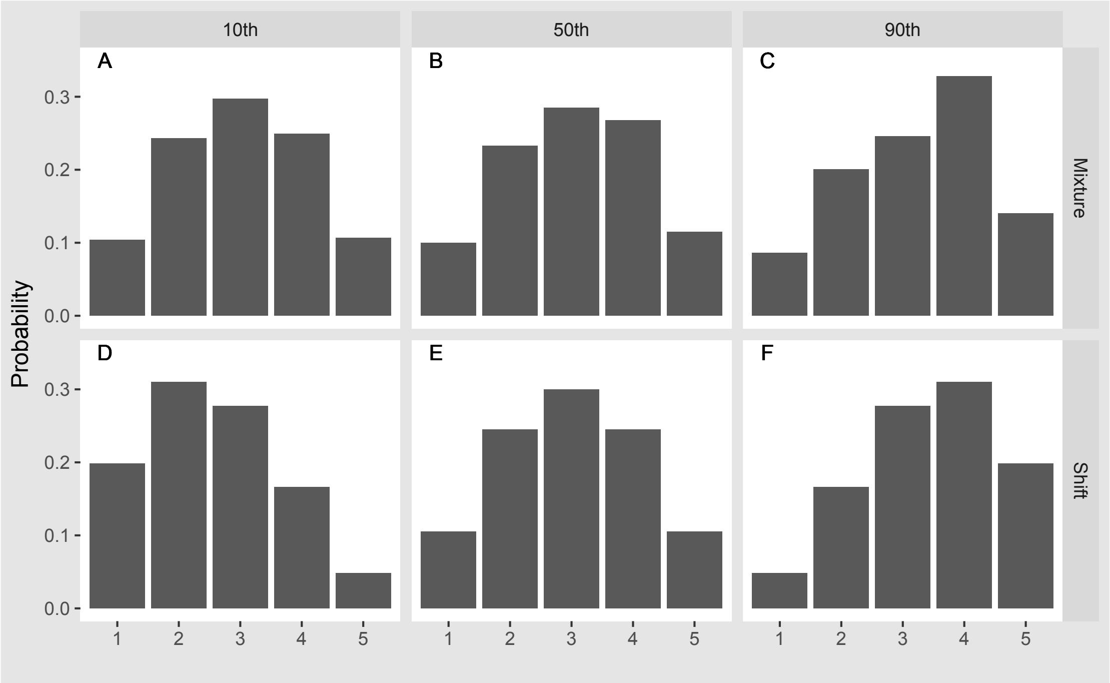


FIGURE 4: The effect of a mixture versus a shift model on the predicted response distributions of a hypothetical 5-point item. Displayed are the probabilities for respondents at the 10th, 50th, and 90th percentile of the ARS distribution.

Ferrando, Morales-Vives, & Lorenzo-Seva, 2016; Johnson & Bolt, 2010; Kam & Zhou, 2015; Maydeu-Olivares & Coffman, 2006). However, the basic idea of two additive, compensatory components (i.e., θ^*_{ti} and θ^*_{ai}) is common to all these approaches.

It is worth noting the following theoretical implications of this shift model. First, as illustrated in Figure 4, the model implies that acquiescence and disacquiescence are opposite poles of a single dimension. Hence, high ARS-values predict a shift towards agreement (Figure 4F), whereas low ARS-values predict a shift towards disagreement (Figure 4D). Moreover, intermediate θ^*_{ai}-values imply the absence of an ARS effect (Figure 4E). Second, the model is a *compensatory model* because high ARS-levels can be compensated for by low target-trait levels. Third, due to this compensatory nature of the shift model, a (moderate) ARS effect may result in a shift from a *strongly disagree*-response to a *disagree*-response. In such a case, however, ARS actually predicts disagreement, despite the theoretical definition of acquiescence as the tendency to prefer agree responses.

Contrary to this shift model, the novel Acquiescence Model is a *mixture model* of acquiescence. Agreement with an item may emerge from two distinct cognitive processes, namely, either from the target trait or from ARS (see Figure 3). The observable distribution of response frequencies is not a single distribution shifted by ARS, but a mixture of two underlying distributions with mixing probabilities

a_{ij} and $1 - a_{ij}$ (see Equation 11). Figure 4 illustrates three implications, in which this mixture model qualitatively differs from the shift model: First, the opposite pole of ARS is the absence of ARS and not disacquiescence. That is, probabilities a_{ij} close to 0 imply the absence of an ARS effect (Figure 4A). In such a case, the Acquiescence Model (Figure 3) reduces to the Böckenholt Model (Figure 2) because the ARS-branch is never reached. Second, the Acquiescence Model is *non-compensatory* (e.g., Babcock, 2011) in nature. Substantively, high levels of acquiescence result in entering the ARS-branch. In this case, the lower branch of the tree and especially the target-trait parameter θ_t do not affect the predicted response probabilities at all and cannot compensate for high ARS levels. Third, an increase in a_{ij} increases the probabilities for the two *agree*-categories and decreases the probabilities for the three non-agree categories. This means that an ARS effect may shift disagreement (predicted by t_{ij}) to agreement, but—in contrast to the shift model—a shift from a *strongly disagree*-response to a *disagree*-response is impossible.

Besides these substantive differences, the shift and the mixture approach also share several properties. In both models, higher ARS-levels increase the probability of affirmative responses. Furthermore, high ARS may lead to agreement with both regular and reverse-coded items in both models.

However, the question remains which model is more appropriate given their diverging implications. First, definitions of ARS are not very precise and can be interpreted in either way. Operationalizations of ARS differ: For example, Couch and Keniston (1960) computed the mean across both regular and reverse-coded items as a measure of ARS—which is in line with a shift model; in contrast, Billiet and McClendon (2000) counted the number of affirmative responses—which is in line with a mixture model. Second, from the perspective of a shift model, measures of acquiescence and disacquiescence should show strong, negative correlations. However, values of .28 (Weijters et al., 2010b) or $-.16$ (Baumgartner & Steenkamp, 2001) are not consistent with this prediction. Third, the Spinozan procedure discussed above is only compatible with the non-compensatory mixture approach: Agreement may result either from deliberate assessment in the second stage or from initial acceptance in the first stage that is neither confirmed nor corrected because of premature output. Contrarily, the shift account is neither compatible with the Spinozan procedure nor are the authors aware of a theoretical account that explains the shift approach. In summary, this hints at the possibility that the novel mixture account might fare better both theoretically and empirically. Furthermore, the development of appropriate statistical models is a prerequisite for comparing the different accounts of ARS in the first place.

Bayesian IRT: Priors and Estimation

We adapted the hierarchal priors and weakly informative hyperpriors proposed by Matzke et al. (2015), which are similar to those in standard Bayesian IRT models with random person and random item effects (see also Fox, 2010). For the Böckenholt Model, we checked that these priors resulted in parameter estimates closely resembling those based on the maximum-likelihood analysis proposed by Böckenholt (2012). For the person parameters, we assume a centered, multivariate normal distribution,

$$\boldsymbol{\theta}_i \sim \text{Multivariate-Normal}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{18}$$

with a covariance matrix $\boldsymbol{\Sigma}$ estimated from the data. In contrast, the item parameters have independent, univariate normal priors. To allow for the possibility to define different hyperpriors for response styles and the target trait(s) (indexed by $p = (e, m, a, t_1, \ldots, t_K)$), the item parameters β are partitioned into an additive combination of mean μ and centered differences δ:

$$\beta_{pj} = \mu_p + \delta_{pj}. \tag{19}$$

Note that such a decomposition also improves convergence when estimating the parameters (Matzke et al., 2015). Based on this parameterization, the following prior and hyperprior distributions were used:

$$\begin{aligned}
\mu_p &\sim \text{Truncated Normal}(0, 1, -5, 5) & (20)\\
\beta_{e^*} &\sim \text{Truncated Normal}(0, 1, -5, 5) & (21)\\
\delta_{pj} &\sim \text{Truncated Normal}(0, \sigma^2_{\beta_p}, -5, 5) & (22)\\
\sigma^2_{\beta_p} &\sim \text{Inverse-Gamma}(1, 1) & (23)\\
\boldsymbol{\Sigma} &\sim \text{Scaled Inverse-Wishart}(\boldsymbol{I}_P, df = P + 1, \tau_p) & (24)\\
\tau_p &\sim \text{Uniform}(0, 100). & (25)
\end{aligned}$$

The priors for the item parameters μ_p, β_{e^*}, and δ_{pj} were truncated to aid faster convergence. Given that latent probit values larger than 5 result in negligible response probabilities smaller than $3 \cdot 10^{-7}$, this does not constrain or inform parameter estimation substantially (but this restriction can also be dropped). The inverse-Wishart prior for the covariance matrix $\boldsymbol{\Sigma}$ in Equation 24 (parameterized by $P + 1$ degrees of freedom and the P-dimensional identity matrix $\boldsymbol{I}_P$) implies marginal uniform priors on the correlations. Moreover, the scaled version of the inverse-Wishart prior with scale parameters τ_p maintains this property but is less restrictive with respect to the variances in $\mathbf{\Sigma}$ (Gelman et al., 2014, p. 74).

Because an analytical solution for the full posterior is not available, the model is estimated by approximating the posterior distribution by Markov chain Monte Carlo (MCMC) sampling using JAGS (Denwood, 2016; Plummer, 2003), a popular software for Gibbs sampling. To cross-check our results, we also implemented the model in Stan (Carpenter et al., 2017), a more recent software package that draws posterior samples based on adaptive Hamiltonian Monte Carlo (Hoffman & Gelman, 2014), a more efficient sampling scheme that often reduces auto-correlation. R (R Core Team, 2016) was used as a front-end in both cases, and an R package for parameter estimation is available from https://github.com/hplieninger/mpt2irt/.

Simulation Studies

We performed two simulation studies (a) to investigate the recovery of core parameters of the Acquiescence Model and (b) to compare the Böckenholt Model and the Acquiescence Model when fit to data generated from each of both.

The simulations were summarized using the posterior medians $\hat{\pi}_p$ as estimates for the true parameters π_p (where π_p stands for the person and item IRT parameters θ_{pi} or β_{pj}). In each replication and for each parameter π_p, three measures were calculated across persons or items, namely the correlation $r_{\hat{\pi}_p,\pi_p}$, the mean bias (i.e., $\text{Mean}(\hat{\pi}_p - \pi_p)$), and the RMSE (i.e., $\sqrt{\text{Mean}(\hat{\pi}_p - \pi_p)^2}$). Below, we report summaries of these three measures across replications.

Study 1: Parameter Recovery

Method

Data were generated from the Acquiescence Model for 1,000 persons. A condition with 20 and a condition with 40 items was realized with half of the items being reverse-coded. In each replication, the person parameters were drawn from a centered multivariate normal distribution, $\mathbf{\Theta} \sim \text{MVN}(\mathbf{0}, \mathbf{\Sigma})$. The variances in $\mathbf{\Sigma}$ were set to $\sigma^2_{\theta_m} = \sigma^2_{\theta_e} = \sigma^2_{\theta_a} = 0.33$ and $\sigma^2_{\theta_t} = 1.00$ mirroring the fact that content-related variance is usually larger than response-style-related variance in empirical data (e.g., Billiet & McClendon, 2000). The covariances in $\mathbf{\Sigma}$ were drawn from a

Wishart distribution with $df = 50$ and the scale matrix

$$\mathbf{\Sigma}^* = \begin{bmatrix} 1 & -.2 & 0 & 0 \\ -.2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{26}$$

which mirrors the empirical finding of a negative correlation between MRS and ERS (with 90 % of the simulated correlations in the range $[-.41, 0.04]$), but small correlations otherwise (90 % in the interval $[-.23, .23]$). Furthermore, the item parameters in each replication were drawn from independent truncated normal distributions with $\mu_{\beta_m} = \mu_{\beta_e} = \mu_{\beta_{e*}} = \Phi^{-1}(.70)$, $\mu_{\beta_a} = \Phi^{-1}(.95)$, and $\mu_{\beta_t} = \Phi^{-1}(.50)$, and with $\sigma^2_{\beta_m} = \sigma^2_{\beta_e} = \sigma^2_{\beta_a} = \sigma^2_{\beta_{e*}} = 0.10$ and $\sigma^2_{\beta_t} = 0.50$.[5] This implies, for example, that the expected probabilities for an average person with $\boldsymbol{\theta}_i = \mathbf{0}$ were .05 versus .95 for an ARS versus non-ARS response, respectively, and .30 versus .70 for an MRS versus non-MRS response, respectively (conditional on a non-ARS response). Furthermore, the item parameters β_{tj} of the target trait were simulated with a larger variance than those for the three response style processes.

We generated 200 data sets for each of the two conditions with 20 and 40 items, respectively, and fit the Acquiescence Model using Stan. Posterior samples were obtained from three independent chains with 1,000 iterations each, of which the first 500 were discarded. In preliminary analyses, these values were fine-tuned to balance computation time and precision.

Results

Concerning the correlations of data-generating and fitted parameters, recovery was better for the item than for the person parameters, for obvious reasons: The former were informed by 1,000 persons whereas the latter were informed by only 20 and 40 items, respectively (see Figure 5A). Recovery generally improved when using more items, which is an indication of the consistency of the estimation procedure. With respect to the β-parameters, recovery was best for the β-parameters for MRS and ERS, because these two processes can directly be inferred from observed responses. The tree in Figure 3 illustrates this property of the Acquiescence Model, that midpoint and extreme responses uniquely emerge from MRS and ERS, respectively. In contrast, affirmative responses are either due to the target trait t_{ij} or due to acquiescence a_{ij}, which diminishes the precision of the estimates for β_t and β_a. Moreover, the generated ARS item parameters β_a were much larger than

[5] These values are partly based on theoretical considerations (e.g., the ARS prevalence should be very low) and partly based on preliminary analyses of empirical data sets.

the other parameters, mirroring a low prevalence of acquiescent behavior, and this additionally reduces the precision of parameter estimates, which is further illustrated in Appendix A. With respect to the θ-parameters, a similar pattern as for the β-parameters was observed, with the exception that recovery of θ_t was best due to the fact that the variance of this dimension was larger than that of the three other dimensions.

Concerning the mean bias for MRS, ERS, and the target trait, the estimates for both the item parameters β_{pj} and the person parameters θ_{pi} were unbiased (see Figure 5B). The ARS parameters, however, were overestimated due to the low prevalence of ARS (see also Appendix A). For the β_a-parameters, this guards against a type I error of incorrectly classifying an item as suspicious (i.e., being susceptible to ARS) at the cost of statistical power. Bias was less severe for the θ_a-parameters, and this bias was in particular caused by upwards-shrinkage towards zero for persons low on ARS (because it is hard to tell from only a few items whether such a person has a θ_a-parameter of, say, -0.7 or -0.5). The results for RMSE mirrored those of the correlations reported above. RMSE was generally smaller for the item parameters β than for the person parameters θ, and generally smaller with 40 compared to 20 items (see Figure 5C). Aside from the core parameters for MRS, ERS, ARS, and the target trait, recovery of the single parameter β_{e*} that measures extreme responding conditional on ARS was comparable to that of the β_e-parameters with respect to correlation, bias, and RMSE.

Study 2: Model Recovery

Method

In the second study, 200 data sets were generated from the Böckenholt Model and 200 from the Acquiescence Model. The Böckenholt as well as the Acquiescence Model were fit to each data set. Each data set included 250 persons and 20 items (half of which were reverse-coded). The data-generation procedure was identical to that of Study 1 with the obvious exception that the a_{ij}-parameter is absent in the Böckenholt Model (and the covariance matrix $\boldsymbol{\Sigma}$ reduces to three dimensions).

The models were fit using JAGS which facilitates the computation of the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & Van Der Linde, 2002), a model-selection criterion that trades-off goodness of fit (i.e., minus the expected deviance) against model complexity (i.e., the effective number of free parameters). To select a model, DIC is computed for each of the competing models and the one with the smallest DIC value is selected. Again, we sampled three chains each with 500 retained iterations, but the computations in JAGS required

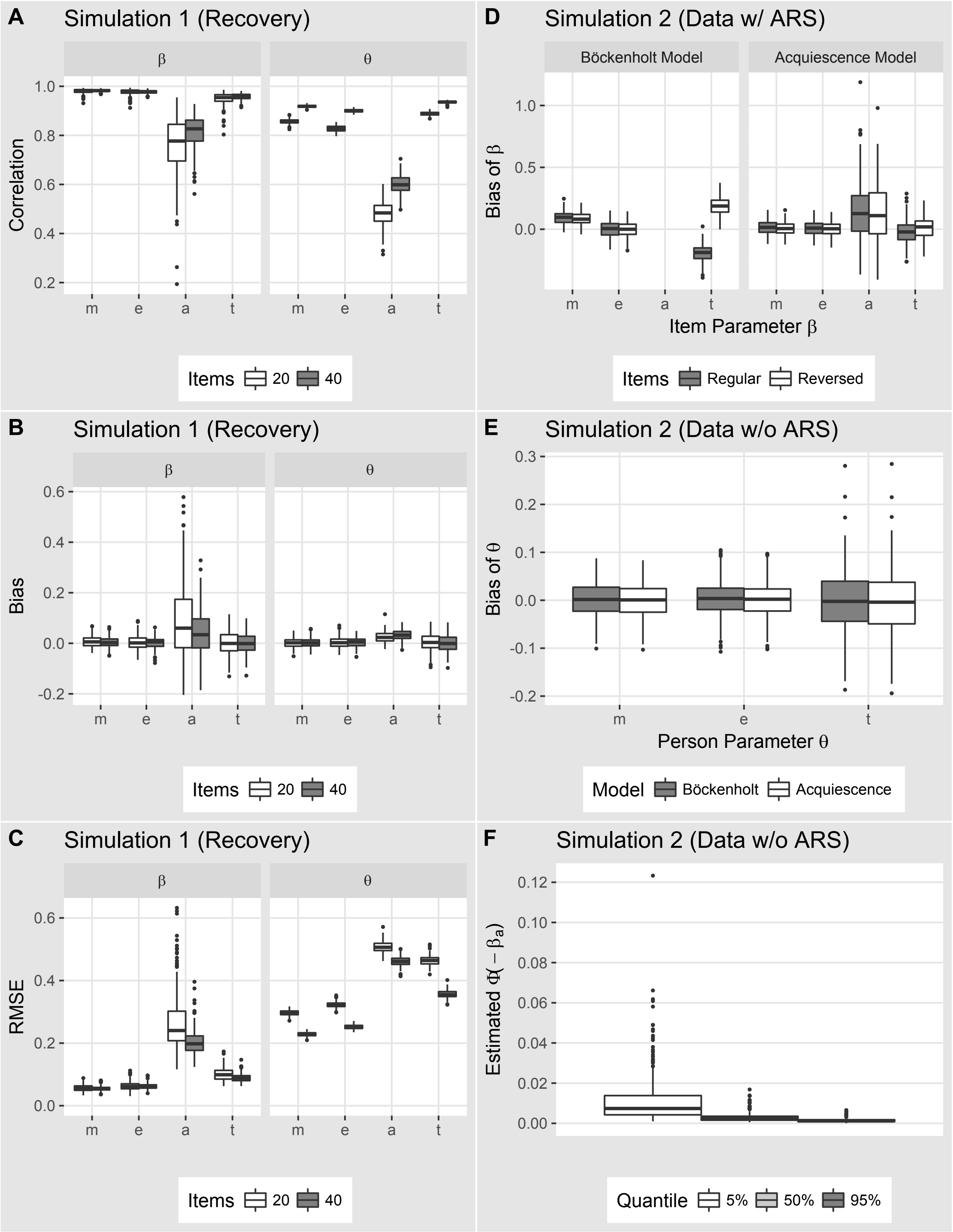


FIGURE 5: Boxplots in panels A to C display the results of Study 1, where the Acquiescence Model was the data-generating and the fitted model. Boxplots in panels D to F display the results of Study 2, where the data-generating model was either the Acquiescence Model (Panel D) or the Böckenholt Model (Panels E and F).

a longer burn-in phase of 1,000 iterations and a thinning factor of 10 (i.e., the number of iterations was actually 10 times larger, but only every 10th iteration was retained).

Results

At first, we consider the condition in which data were generated from the Acquiescence Model. As expected, DIC was smaller for the Acquiescence compared to the Böckenholt Model in 82 % of the cases, indicating that the two models can, with a realistic set-up of 250 persons and 20 item, in principle be discriminated using this model-selection criterion. The preference of DIC towards the Acquiescence Model increased in data sets in which acquiescence was more prevalent: ΔDIC was lower (i.e., more favorable of the Acquiescence Model) the larger the ARS variance $\sigma^2_{\theta_a}$ was ($r = -.21$), and ΔDIC was lower the lower the mean of the β_a-parameters was (thus eliciting ARS responses more easily; $r = .31$).

A closer look at the estimated parameters shows that the item parameters, compared to the person parameters, were more affected when fitting the Böckenholt Model instead of the data-generating Acquiescence Model, an effect that was largest for the target-trait difficulties β_t. Average recovery of these parameters was considerably better when fitting the correctly specified Acquiescence Model ($r_{\hat{\beta}_t,\beta_t} = .86$, $RMSE = 0.16$) instead of the misspecified Böckenholt Model ($r_{\hat{\beta}_t,\beta_t} = .69$, $RMSE = 0.25$). This was due to the fact that the Böckenholt Model does not take into account the confounding of the target trait t_{ij} and ARS a_{ij}; thus, the β_t-parameters were biased, namely, underestimated for regular items and overestimated for reverse-coded items (see Figure 5D). Furthermore, the β_m-parameters were biased if ARS was not taken into account. The person parameters estimated from the Böckenholt Model were—even though inferior to those from the Acquiescence Model—less affected than the item parameters, and therefore do not receive further discussion herein.

Next, we consider the condition in which the data were generated using the Böckenholt Model, that is, without acquiescence. Since this model is nested within the Acquiescence Model, the simulated data are still compatible with the latter model. However, model selection should prefer the Böckenholt Model because of its smaller complexity due to the absence of any (here superfluous) ARS parameters. This was indeed the case, DIC was smaller for the Böckenholt compared to the Acquiescence Model in 90 % of the cases indicating that this criterion again allowed to discriminate between the two models. With respect to recovery of the person parameters, the two models performed equally well, and the Böckenholt Model as well as the overly complex, misspecified Acquiescence Model resulted in unbiased estimates (see Figure 5E). However, the Böckenholt Model was slightly more accurate in estimating the item parameters, especially the target-trait difficulties β_t.

Of special interest were the estimates for the ARS-parameters when fitting the unnecessarily complex Acquiescence Model to data generated without acquiescence. Essentially, the simulation indicated that the Acquiescence Model empirically reduced to the Böckenholt Model in many respects. First, the estimates for the item parameters β_a were very large mirroring a very low prevalence of ARS. Figure 5F illustrates that, for an average person, the probability to give an ARS response to the item at the median of the 20 items was below 1 % in almost all replications. Second, the variance $\sigma^2_{\theta_a}$ was estimated to be .14 on average, whereas the data-generating variances of MRS and ERS ($\sigma^2_{\theta_m} = \sigma^2_{\theta_e} = .33$) were properly recovered in both models. Taken together, if the more complex Acquiescence Model is fit to a data set with absolutely no acquiescence, DIC is likely to indicate a preference for the more parsimonious Böckenholt model and estimates for the ARS parameters β_a and $\sigma^2_{\theta_a}$ will lead to the correct conclusion that responses were not affected by ARS. In such a case, the estimates for the remaining parameters (especially the item parameters) from the Böckenholt Model are expected to be slightly more precise and are thus preferred when drawing substantive conclusions.

Empirical Study

In the following example, we demonstrate the application of the proposed model with empirical data. First, we fit different response style models for a single content domain followed by models comprising six target traits ($\theta_{t_1}, \ldots, \theta_{t_6}$) for a six-dimensional questionnaire.

Method

We re-analyzed data by Moshagen, Hilbig, and Zettler (2014), who investigated the factorial structure and psychometric quality of the German version of the HEXACO personality inventory. Here, we focus on the revised HEXACO personality inventory (HEXACO-PI-R; K. Lee & Ashton, 2016) that includes 96 five-point items (*strongly disagree* to *strongly agree*). This questionnaire is comprised of six scales—namely, honesty-humility (H), emotionality (E), extraversion (X), agreeableness (A), conscientiousness (C), and openness to experience (O)—with the corresponding items presented in alternating order. Within each scale, between seven and ten of the 16 items were reverse-coded (i.e., disagree responses are indicative of a high trait-level) making up a total of 48 reverse-coded items (i.e., 50 % of the items). Our reanalysis is based on the second sample of Moshagen et al. (2014) that includes 1,012 university students, of which we drew a random subsample of 500 respondents. Note that the remaining persons were used for cross-validating the estimated item parameters (see below).

Two response style models, the Böckenholt and the Acquiescence Model, were fit using Stan (Carpenter et al., 2017) by sampling six chains with 1,500 iterations each, of which the first 500 were discarded (i.e., 6,000 retained iterations in total). The number of retained iterations was quadrupled in comparison to the simulation studies, because a good approximation of the posterior was even more important herein. Convergence of the sampling procedure is illustrated in Appendix B, for example, by means of trace plots. Again, the posterior distributions were summarized using their medians as estimates as well as 95 % posterior intervals reported in brackets. Note that, for ease of interpretation, the raw item parameters β were transformed to probabilities $\Phi(0-\beta)$, that is, the probability for an average person with $\boldsymbol{\theta}_i = \mathbf{0}$ to pass the item's threshold. For example, an item parameter $\beta_m = 1$ can be expressed as $\Phi(-1) = .16$: Thus, the probability for an MRS response is 16 % for an average person.

Results

Response style models for a single content domain

We first applied different response style models to a single content domain, namely, honesty-humility, which refers to individual differences in treating "others fairly even when one could successfully exploit them" (K. Lee & Ashton, 2016, p. 2). The scale is comprised of six regular and 10 reverse-coded items. First, we compared the novel Bayesian implementation of the Böckenholt Model to the previously used maximum-likelihood estimation that is based on binary pseudoitems (Böckenholt, 2012; De Boeck & Partchev, 2012). Both the item and the person parameters were virtually identical with correlations above .99 and a mean bias of almost zero. Second, the Acquiescence Model was fit to the data, which resulted in a substantial improvement with DIC of 20,167 compared to 20,296 for the Böckenholt Model.[6]

With respect to the item parameters, most variability was observed for β_t (target-trait difficulty) even though most items were rather easy (i.e., easily eliciting responses indicative of high honesty-humility) with a mean $\Phi(-\mu_{\beta_t})$ of .87 $[.77, .93]$. The probability-transformed MRS item parameters $\Phi(-\beta_m)$ ranged from .15 to .34 with a mean of .22 $[.16, .29]$ indicating that (given a non-ARS response) the middle category was on average chosen with a probability of 22 %. Next, the

[6] The DIC was always estimated separately using JAGS with a thinning factor of 10.

ERS difficulties $\Phi(-\beta_e)$ ranged from .14 to .48 with a mean of .30 $[.22, .39]$ indicating that—when choosing between an extreme and a moderate category—an extreme response was on average chosen with a probability of 30 %. Of most interest, the novel ARS difficulties $\Phi(-\beta_a)$ were estimated to be considerably low ranging from .01 to .16 with a mean of .04 $[.02, .08]$. Substantially, this implies that an ARS response was on average expected in only 4 % of the cases. This prevalence might seem rather small at first sight, but it was expected given the model definition. Essentially, ARS responses are assumed to be independent of content-related response processes and should therefore occur infrequently when using both dependable samples and solid psychometric questionnaires. However, a few items showed higher levels of acquiescence, an observation that is discussed further in the analysis below.

Besides the item parameters, the person parameters allow for additional insights. As expected, response style variance was smaller than target-trait variance with estimates of $\sigma^2_{\theta_m} = 0.21$ $[0.16, 0.26]$, $\sigma^2_{\theta_e} = .49$ $[0.41, 0.60]$, $\sigma^2_{\theta_a} = 0.32$ $[0.17, 0.51]$, and $\sigma^2_{\theta_t} = 1.13$ $[0.88, 1.46]$. Note that the variance of acquiescence, which is ignored in the Böckenholt Model, was estimated to be larger than that of MRS, thereby indicating the importance of ARS. Correlations between different response styles were estimated to be rather small with the exception of MRS and ERS, which correlated negatively at $-.53$ $[-.64, -.41]$, which is a substantively plausible finding given that extreme responses are likely to be accompanied by less midpoint responses.

Response style models for all six HEXACO scales

The target trait(s) and response styles are in general more easily disentangled when using content-heterogeneous items as found in multidimensional questionnaires, thereby facilitating the detection and estimation of response styles (e.g., Khorramdel & von Davier, 2014; Weijters et al., 2010b). The same holds for the Böckenholt and the Acquiescence Model, which we fit to multiple domains simultaneously with the constraint that the response style parameters θ_m and θ_e (and θ_a in the Acquiescence Model) are identical across all items. Thereby, precision of the corresponding estimates is expected to increase considerably, given that only content-related parameters are allowed to vary across different domains. However, a shortcoming of standard estimation techniques (e.g., the EM algorithm; Dempster, Laird, & Rubin, 1977) is that such high-dimensional models become computationally intractable when the number of dimensions becomes large, say larger than four (e.g., Fox, 2010). However, this limitation does not apply to Bayesian implementations. Therefore, we were able to estimate a 9-dimensional version of the Acquiescence Model comprised of six target traits and three response styles.

In terms of fit, the Acquiescence Model was superior compared to the Böckenholt Model with DIC values of 122,348 and 123,024, respectively, indicating the importance of taking ARS into account. The estimated item parameters of the Acquiescence Model are displayed in Figure 6. Across the six HEXACO scales, the content-related item parameters were most variable (with variances ranging from $\sigma^2_{\beta_\text{H}} = 0.51$ to $\sigma^2_{\beta_\text{A}} = 1.16$) and rather easy (with means ranging from $\Phi(-\mu_{\beta_\text{A}}) = .51$ to $\Phi(-\mu_{\beta_\text{H}}) = .87$). In contrast, the MRS and ERS parameters were much more homogeneous with $\sigma^2_{\beta_m} = 0.07$ $[0.06, 0.10]$ and $\sigma^2_{\beta_e} = 0.13$ $[0.10, 0.18]$. Moreover, the thresholds for these response styles were rather high accompanied by low mean probabilities of $\Phi(-\mu_{\beta_m}) = .25$ $[.23, .27]$ and $\Phi(-\mu_{\beta_e}) = .26$ $[.23, .29]$. Substantively, this implies that midpoint and extreme responses were on average given with conditional probabilities of 25 % and 26 %, respectively. The ARS parameters β_a, which were of particular interest here, were rather difficult with $\Phi(-\mu_{\beta_a}) = .03$ $[.02, .04]$ indicating that ARS responses were unlikely for most of the items. However, a few items stood out, for example, $\Phi(-\beta_{a,67}) = .14$ $[.09, .19]$. For this item, a person with an average ARS-level (i.e., $\theta_{ai} = 0$) has a probability of 14 % of agreeing with this item irrespective of his or her standing on the target trait $\theta_{t_k i}$. For a person with an ARS-level θ_{ai} one standard deviation above or below the mean, the probability of an ARS-response changes to 25 % and 7 %, respectively. The probability corresponding to parameter β_{e^*} (i.e., the ERS-difficulty conditional on ARS) was rather low with $\Phi(-\beta_{e^*}) = .03$ $[.02, .05]$, thereby indicating that the *agree*-category was preferred over the *strongly agree*-category in the ARS-branch.

Regarding the person parameters, most variability was observed with respect to the target trait (with variances ranging from $\sigma^2_{\theta_\text{O}} = .65$ $[0.53, 0.81]$ to $\sigma^2_{\theta_\text{H}} = 1.20$ $[0.96, 1.52]$), compared to smaller variances for the response-style-related processes (i.e., $\sigma^2_{\theta_m} = 0.06$ $[0.05, 0.08]$, $\sigma^2_{\theta_e} = 0.26$ $[0.23, 0.31]$, and $\sigma^2_{\theta_a} = 0.16$ $[0.11, 0.23]$). Similarly as in the single-domain analysis, MRS was negatively correlated with ERS and with ARS, in contrast to a positive correlation between ERS and ARS (see Table 1). Importantly, the content–style correlations were rather small with a mean absolute correlation of .09 (see Table 1). Larger values were observed for the relationship of ARS with both honesty-humility ($r = -.36$ $[-.51, -.20]$) and conscientiousness ($r = -.24$ $[-.40, -.08]$). Even though research on the relationship between ARS and the HEXACO traits is sparse, these correlations seem plausible at face value: Pretentious, hypocritical as well as sloppy, negligent persons (Ashton & Lee, 2007) were more prone to ARS compared to sincere as well as careful persons.

Apart from that, the intercorrelations of the six target traits were in a range

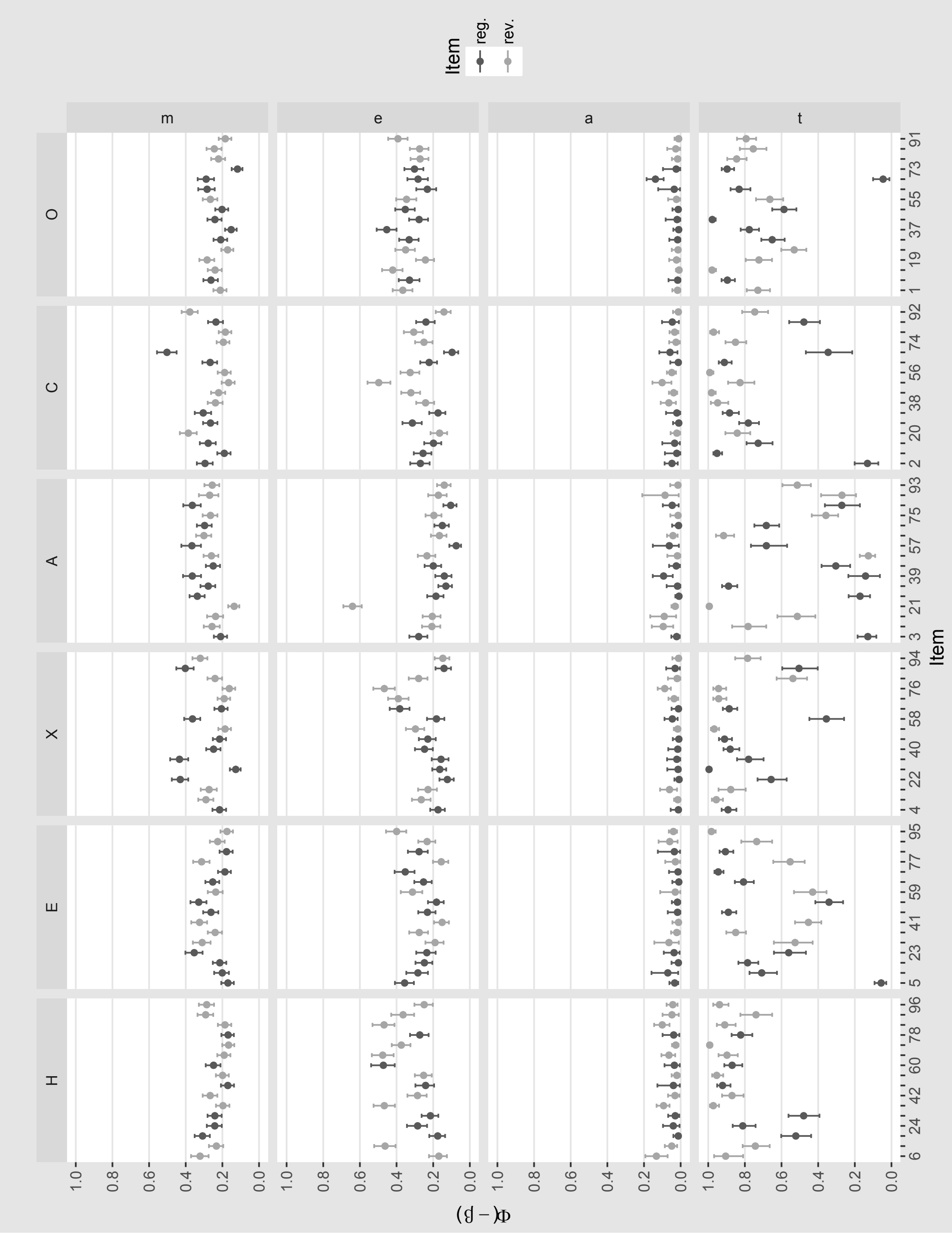


FIGURE 6: Posterior median and 95 % posterior intervals for the probability-transformed item parameters as a function of (horizontally) the psychological process, (vertically) the corresponding HEXACO scale (Honesty-humility, Emotionality, eXtraversion, Agreeableness, Conscientiousness, and Openness), and (black vs. gray) item coding direction.

Table 1: Estimated Latent Correlations and Variances of the Acquiescence Model for all 96 HEXACO Items

	Response styles			Personality traits					
	MRS	ERS	ARS	H	E	X	A	C	O
MRS	0.06	[−.49, −.30]	[−.45, −.14]	[−.20, .04]	[−.08, .15]	[−.28, −.05]	[−.09, .14]	[−.04, .20]	[−.26, −.03]
ERS	−.40	0.26	[.27, .53]	[−.14, .09]	[−.06, .15]	[−.12, .09]	[−.12, .09]	[−.14, .08]	[−.08, .14]
ARS	−.30	.41	0.16	[−.51, −.20]	[−.23, .08]	[−.29, .04]	[−.27, .04]	[−.40, −.08]	[−.21, .12]
H	−.08	−.02	−.36	1.20	[−.05, .18]	[−.15, .09]	[.23, .45]	[−.02, .23]	[.04, .28]
E	.03	.05	−.07	.07	0.77	[−.14, .10]	[−.34, −.12]	[−.07, .17]	[−.08, .16]
X	−.17	−.01	−.13	−.03	−.02	1.10	[−.11, .12]	[.07, .31]	[.03, .28]
A	.03	−.02	−.12	.35	−.23	.01	0.96	[−.15, .10]	[−.07, .17]
C	.08	−.03	−.24	.11	.05	.19	−.03	0.81	[−.16, .09]
O	−.15	.03	−.05	.16	.05	.16	.06	−.03	0.65

Note. Values on the diagonal are variances, values on the lower triangular matrix are correlations, and values on the upper triangular matrix are 95 % posterior intervals of correlations.

from $-.23$ to .35 (see Table 1). Importantly, these correlations hardly differed from those estimated from a standard ordinal model, namely, a Steps model[7] for six correlated target traits. The mean of the absolute differences between these two models was 0.019 and the largest difference was 0.046. This finding has two implications. First, controlling for response styles did not change the substantive conclusions about the intercorrelations in our data, which is in line with previous work (e.g., Plieninger, 2016). Second, the model entails, figuratively speaking, a crude dichotomization of the items into agreement versus disagreement. This, however, did not remove much substantive variance from the data at least concerning the intercorrelations.

When comparing the parameter estimates of the multidimensional Acquiescence Model with those from the single-domain version for honesty-humility above, the item parameters were almost identical with $r > .99$. While a similarly high correlation was observed for the person parameters θ_t measuring honesty-humility ($r = .98$), the response-style person-parameters showed smaller, but still high correlations between the two model version (.48 for MRS and .70 for both ERS and ARS). In line with our simulation studies, this illustrates that person parameters for response styles are estimated more precisely when using more items (96 vs. 16). In addition, the multidimensional model in principle benefits from the fact that a questionnaire with multiple content domains allows for a better discrimination between target traits and response styles.

Posterior predictive checks Model fit was further evaluated by means of posterior predictive checks (e.g., Gelman et al., 2014; M. D. Lee & Wagenmakers, 2013), that is, by assessing the discrepancy between the actually observed response frequencies and response frequencies generated based on the posterior samples. For this purpose, we simulated individual responses for each item using a subsample of the parameter vectors from the joint posterior distribution (namely, every 100th iteration in each of the six chains for every person and for every item). These posterior-predictive response frequencies were then aggregated across persons mirroring (for each item) the response distribution implied by the fitted model. Across these 600 replications, we compared the predicted 68 % and 95 % posterior intervals against the observed response distribution to test whether the model accurately accounted for the observed data. This was done separately for the Acquiescence Model, the Böckenholt Model, and a Steps Model that does not account for response styles at all.

[7] The Steps model (Tutz, 1990; Verhelst, Glas, & de Vries, 1997) is an ordinal IRT model without response styles and was proposed as an alternative to, for example, the partial credit or the graded response model. It serves as natural comparison model herein, because it is also based on a tree-structure (De Boeck & Partchev, 2012).

Posterior predictive checks are illustrated in Figure 7A, namely, for respondents above the 90th percentile of the ARS distribution and for the three items most susceptible to ARS (i.e., low β_{aj}). If acquiescence affected response behavior, the Acquiescence Model should outperform the two other models because it explicitly accounts for ARS. Figure 7A shows that this is indeed the case. The Acquiescence Model was superior to the two competitors especially in predicting response frequencies in Category 4 (i.e., *agree*). This holds also for all 96 items: The coverage rate (i..e, percentage of 95 % posterior intervals covering the observed frequencies) equaled 95 % for the Acquiescence Model, 90 % for the Böckenholt Model, and 83 % for the Steps Model.

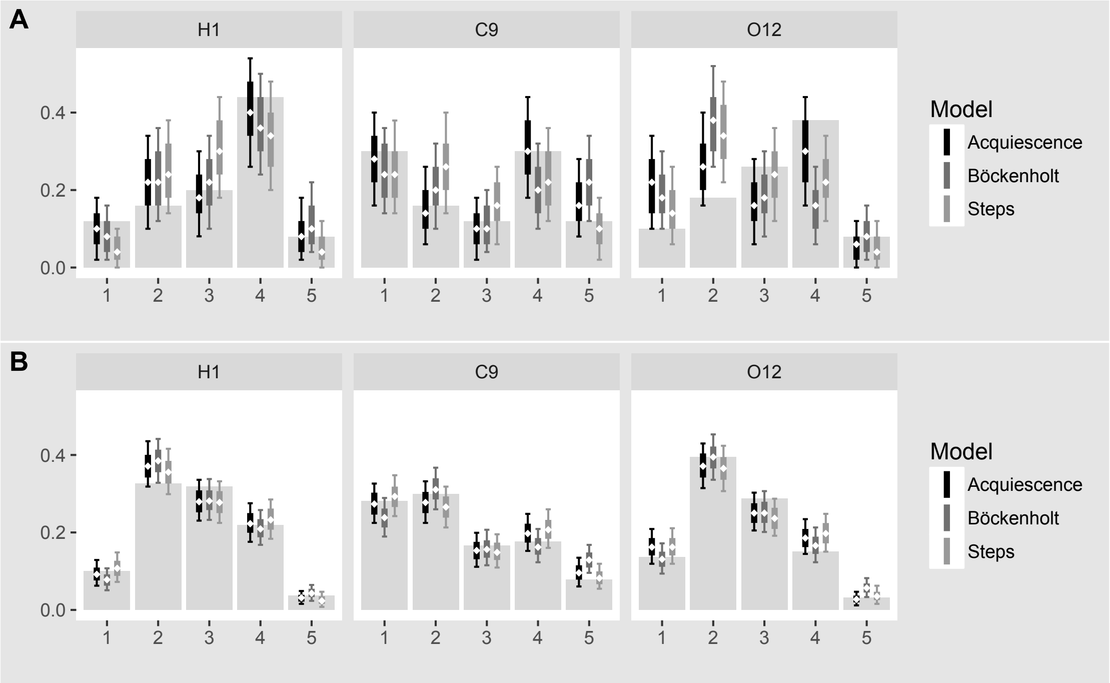


FIGURE 7: Posterior predictive checks for three selected items. The thin (fat) error bars represent 95 % (68 %) posterior intervals from the posterior-predictive distribution of three different models. Panel A displays the observed frequencies of the subset of persons from the original sample with ARS estimates above the 90th percentile. Panel B displays the observed frequencies of all persons in the cross-validation sample and the out-of-sample predictions for hypothetical respondents.

We also performed a cross-validation check. In principle, it is possible that the proposed model leads to overfitting and only provides a good approximation of the fitted data, but that the model performs poorly with respect to predicting new data. Therefore, the quality of out-of-sample predictions was investigated using the remaining 512 respondents from the sample of Moshagen et al. (2014) as a cross-validation data set. The bars in Figure 7B show the observed frequencies for these respondents for the same three items as before, whereas the error bars indicate the predicted data for 512 hypothetical persons sampled from the multivariate posterior distribution of the person parameters. The Acquiescence Model predicted the new, empirical response frequencies very well with all 95 % and most 68 % posterior intervals covering the observed data. Whereas the Steps Model performed comparatively well, the Böckenholt Model provided a worse prediction, especially for the *strongly agree*-category. With respect to all 96 items, the Acquiescence Model with a coverage rate of 92 % again outperformed the Böckenholt Model with 87 % coverage, while the Steps Model performed slightly better with 97 % coverage. In summary, the posterior predictive checks indicated that the Acquiescence Model successfully predicted the fitted data and also new response frequencies thereby corroborating the model-selection results based on DIC above.

Mixture vs. shift model Above, we theoretically compared the Acquiescence Model, which defines ARS as a mixture process, with alternative accounts that model ARS as a shift process. To compare both accounts empirically, we additionally fit a classical confirmatory factor model (e.g., Billiet & McClendon, 2000) to the data and retrieved the resulting person parameters θ^*_{ai} for the ARS factor. Additionally, we computed descriptive proxy variables, which have been proposed as indices for acquiescence, for every person, namely, $A1$, the mean across all items before recoding (e.g., Couch & Keniston, 1960) and $A2$, the number of responses in the affirmative response categories *agree* and *strongly agree* (e.g., Billiet & McClendon, 2000).

Table 2 shows the correlations of these four measures, indicating that the shift model was in close agreement with the mean-index $A1$, much closer than with the count-index $A2$. The opposite pattern was found for the mixture model, which was more strongly related to $A2$ than to $A1$. These correlations thereby reflect the definition of the two indices $A1$ and $A2$, which are descriptively defined as shifts in item means versus changes in the number of affirmative responses, respectively. Concerning the model-based ARS estimates, the medium-sized correlation between the Acquiescence Model and the factor model (i.e., between θ_a and θ^*_a) indicated that there was substantial, but imperfect overlap between the definition of acquiescence in the mixture and the shift model. These results highlight that both models measure distinct albeit related constructs. Moreover, this implies that the definition of acquiescence in terms of a mixture or a shift process has consequences with respect to the appropriate measurement of ARS.

Discussion

We developed a new model for acquiescence, a response style characterized by a preference for affirmative response categories. Inspired by MPT models popular in

Table 2: Correlations of Model-Based and Descriptive Measures of Acquiescence

	Model-based		Descriptive	
	θ_a	θ_a^*	$A1$	$A2$
θ_a	—	.36	.45	.71
θ_a^*		—	.96	.66
$A1$			—	.69

Note. θ_a: Acquiescence Model; θ_a^*: confirmatory factor model; $A1$: mean across all items; $A2$: number of agreements.

cognitive psychology, the new model builds on the work of Böckenholt (2012) and explicitly assumes a theoretically motivated tree-like structure of latent cognitive processes. Extending the original model, the new Acquiescence Model allows to capture not only ERS and MRS, but also ARS. All of these processes are modeled using an IRT approach, namely, by additive person and item effects on the probit scale. Within the proposed model, agreement is conceptualized as a mixture process—stemming either from a high target-trait level or from ARS.

The starting point, the Böckenholt Model in Figure 2 assumes three qualitatively different processes. Whereas MRS directly leads to midpoint responses, the target trait leads to agreement with regular items (and disagreement with reverse-coded items) conditional on non-MRS, and ERS conditionally leads to extreme responses. The Böckenholt Model in specific and IR-tree models in general are characterized by these conditional definitions of response processes similar as in MPT models that are used in cognitive psychology (e.g., Erdfelder et al., 2009; Matzke et al., 2015). In psychometrics, this approach has been proven useful in both methodological and applied work (Böckenholt, 2017; Böckenholt & Meiser, 2017; Jeon & De Boeck, 2016; Khorramdel & von Davier, 2014; Plieninger & Meiser, 2014; Thissen-Roe & Thissen, 2013; Zettler et al., 2016). However, ARS—which is often seen as especially important (e.g., Plieninger, 2016; Rammstedt, Goldberg, & Borg, 2010)—was not included in the Böckenholt Model, a disadvantage compared to alternative response style models (e.g., Johnson & Bolt, 2010). As a remedy, we proposed an extension of the Böckenholt Model, namely, the Acquiescence Model, which assumes an additional ARS process that leads to agreement with an item irrespective of its coding direction or content. Thus, the new model allows to disentangle four different processes—three response styles as well as the target trait.

Each of the four processes is comprised of a person parameter θ, which captures individual differences (e.g., in ARS-responding), and an item parameter β, which captures item specific effects. With respect to the target trait t_{ij}, these item effects correspond to item difficulties. With respect to response styles, these item effects capture difference between items in eliciting a particular response style. It is important to note that those style-related difficulties are not an indication of response styles being dependent on item *content.* Rather, "external" item features such as length, position, or complexity are possible explanations for this heterogeneity (e.g., De Jong et al., 2008), and Figure 6 illustrates that empirical data indeed exhibit such heterogeneity. While previous research mainly focused on person-level covariates of response styles (see Van Vaerenbergh & Thomas, 2013), future research should also consider item-level covariates—and the proposed Acquiescence Model is a well-suited tool for that.

Existing models for ARS (e.g., Billiet & McClendon, 2000; Ferrando et al., 2016) have the advantage that they are easy to interpret and often easy to implement in respective software. Our model shares some similarities with these approaches, and we value these models in general. However, we pose the question whether the existing statistical approaches are fully in line with the concept of acquiescence as currently in use as well as with theoretical reasoning. If ARS is specific to agreement (and relatively independent of disacquiescence), then the proposed mixture approach may be more appropriate. Moreover, if acquiescent responses are indeed due to initial acceptance of an item in the comprehension stage paired with premature output before the assessment stage of the Spinozan procedure (Gilbert, 1991; Knowles & Condon, 1999), then the mixture approach may also be more appropriate. Even though it is to early to answer the posed question, the proposed Acquiescence Model enables an empirical comparison of the two approaches in the first place. Furthermore, it will stimulate a discussion about the theoretical implications of both views in order to gain a deeper understanding of acquiescence.

The proposed model belongs to the general class of hierarchical MPT models (Heck, Arnold, & Arnold, 2017; Klauer, 2010; Matzke et al., 2015), which subsume IR-tree models (Böckenholt, 2012; De Boeck & Partchev, 2012) as special cases, models that have already been proven useful in psychometrics. We believe that hierarchical MPT models provide a general and fruitful framework for future developments and applications following the idea of "cognitive psychometrics" (Riefer et al., 2002). For example, hierarchical MPT models might prove useful in modeling guessing behavior, which also results in a mixture distribution of content-related and content-independent responses similar to ARS in the Acquiescence Model.

Future work may adapt the Acquiescence Model to items with more or less than five categories (see Böckenholt & Meiser, 2017; Plieninger & Meiser, 2014). Moreover, the difference between *agree*- and *strongly agree*-responses conditional on acquiescence may be of further interest. With respect to this "(ARS-)conditional ERS process", we opted for a parsimonious model with only a single item parameter β_{e*}, but more complex models with additional item and/or person parameters can in principle be tested. Furthermore, we chose a restrictive approach and modeled response styles as stable across content domains (but t_{ij} was domain specific). This is in line with previous work (e.g., Danner et al., 2015; Khorramdel & von Davier, 2014; Weijters et al., 2010a) even though others have argued for a different view (e.g., Ferrando, Condon, & Chico, 2004), and it is in principle possible to extend the Acquiescence Model in this direction.

Concerning parameter estimation, the recoding procedure previously used to obtain maximum-likelihood estimates with software for multidimensional IRT models (Böckenholt, 2012; De Boeck & Partchev, 2012) was no longer applicable in the Acquiescence Model. This is due to the mixture structure on the level of item-person combinations, according to which the response probability for affirmative responses compromises two additive, IRT-like parts. As a remedy, we adapted a Bayesian implementation of hierarchical MPT models (Klauer, 2010; Matzke et al., 2015). Besides providing virtually identical estimates as the ML procedure in case of the Böckenholt Model, this gives the researcher great flexibility to fit complex models such as the 9-dimensional Acquiescence Model for the HEXACO data presented above. In two simulation studies, we showed that parameter recovery of the Acquiescence Model was satisfactory. Moreover, fitting the model to data generated without acquiescence (i.e., the standard Böckenholt Model) did not affect conclusions substantially because (a) DIC was likely to select the correct model, and (b) the Acquiescence Model empirically reduced to the Böckenholt Model—that is, the ARS item parameters β_a became extremely large and the ARS person variance $\sigma^2_{\theta_a}$ became relatively small (thereby implying a very low probability of ARS responses). Taken together, these results show that the proposed Acquiescence Model provides a useful generalization of the Böckenholt Model.

The empirical example illustrated that the prevalence of acquiescence was rather low for the German version of the HEXACO-PI-R in the current sample. On the one hand, this finding is reassuring from an assessment perspective, because higher ARS levels might raise validity concerns. On the other hand, the low prevalence of acquiescence makes a precise estimation of the ARS parameters difficult, which was counteracted by using as many as 96 items.

In sum, we proposed the Acquiescence Model to generalize and improve an already successful response style model (Böckenholt, 2012). Thereby, we provide an answer to the question how to account for the empirically relevant and theoretically interesting phenomenon of ARS within the psychologically meaningful tree-like structure of IR-tree models. In modeling ARS as a mixture process, we shed light on the question how to precisely define acquiescent response behavior. To address such theoretical questions in general, we advocate the use of hierarchical MPT models that explicitly account for latent response processes and thereby provide a powerful framework at the interface of psychometrics and cognitive psychology.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/BF02293814

Ashton, M. C. & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, *11*, 150–166. doi:10.1177/1088868306294907

Babcock, B. (2011). Estimating a noncompensatory IRT Model using Metropolis within Gibbs sampling. *Applied Psychological Measurement*, *35*, 317–329. doi:10.1177/0146621610392366

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

Billiet, J. B. & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*, 608–628. doi:10.1207/S15328007SEM0704_5

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

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

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*, 159–181. doi:10.1111/bmsp.12086

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. doi:10.18637/jss.v076.i01

Couch, A. & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*, 151–174. doi:10.1037/h0040372

Danner, D., Aichholzer, J., & Rammstedt, B. (2015). Acquiescence in personality questionnaires: Relevance, domain specificity, and stability. *Journal of Research in Personality*, *57*, 119–130. doi:10.1016/j.jrp.2015.05.004

De Boeck, P. & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. doi:10.18637/jss.v048.c01

De Jong, M. G., Steenkamp, J.-B. E. M., 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*, 104–115. doi:10.1509/jmkr.45.1.104

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38. Retrieved from http://www.jstor.org/stable/2984875

Denwood, M. J. (2016). runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(1), 1–25. doi:10.18637/jss.v071.i09

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models. *Zeitschrift für Psychologie/Journal of Psychology*, *217*, 108–124. doi:10.1027/0044-3409.217.3.108

Ferrando, P. J., Condon, L., & Chico, E. (2004). The convergent validity of acquiescence: An empirical study relating balanced scales and separate acquiescence scales. *Personality and Individual Differences*, *37*, 1331–1340. doi:10.1016/j.paid.2004.01.003

Ferrando, P. J., Morales-Vives, F., & Lorenzo-Seva, U. (2016). Assessing and controlling acquiescent responding when acquiescence and content are related: A comprehensive factor-analytic approach. *Structural Equation Modeling*, *23*, 713–725. doi:10.1080/10705511.2016.1185723

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications.* doi:10.1007/978-1-4419-0742-4

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, *46*, 107–119. doi:10.1037/0003-066X.46.2.107

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

Heck, D. W., Arnold, N. R., & Arnold, D. (2017). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods.* Advance online publication. doi:10.3758/s13428-017-0869-7

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623. Retrieved from http://jmlr.org

Hu, X. & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. doi:10.1007/bf02294263

Hütter, M. & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, *27*, 116–159. doi:10.1080/10463283.2016.1212966

Jeon, M. & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085. doi:10.3758/s13428-015-0631-y

Johnson, T. R. & Bolt, D. M. (2010). On the use of factor-analytic multinomial logit item response models to account for individual differences in response style. *Journal of Educational and Behavioral Statistics*, *35*, 92–114. doi:10.3102/1076998609340529

Kam, C. C. S. & Zhou, M. (2015). Does acquiescence affect individual items consistently? *Educational and Psychological Measurement*, *75*, 764–784. doi:10.1177/0013164414560817

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*, 161–177. doi:10.1080/00273171.2013.866536

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98. doi:10.1007/s11336-009-9141-0

Knowles, E. S. & Condon, C. A. (1999). Why people say "yes": A dual-process theory of acquiescence. *Journal of Personality and Social Psychology*, *77*, 379–386. doi:10.1037/0022-3514.77.2.379

Lee, K. & Ashton, M. C. (2016). Psychometric properties of the HEXACO-100. *Assessment*. Advance online publication. doi:10.1177/1073191116659134

Lee, M. D. & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. New York, NY: Cambridge University.

Mandelbaum, E. (2014). Thinking is believing. *Inquiry*, *57*, 55–96. doi:10.1080/0020174X.2014.858417

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E.-J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika*, *80*, 205–235. doi:10.1007/s11336-013-9374-9

Maydeu-Olivares, A. & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344–362. doi:10.1037/1082-989X.11.4.344

Moshagen, M., Hilbig, B. E., & Zettler, I. (2014). Faktorenstruktur, psychometrische Eigenschaften und Messinvarianz der deutschsprachigen Version des 60-Item HEXACO Persönlichkeitsinventars [Factor structure, psychometric properties, and measurement invariance of the German-language version of the 60-item HEXACO personality inventory]. *Diagnostica*, *60*, 86–97. doi:10.1026/0012-1924/a000112

Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32–53. doi:10.1177/0013164416636655

Plieninger, H. & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899. doi:10.1177/0013164413514998

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2013)*. Vienna, Austria. Retrieved from https://www.r-project.org/conferences/DSC-2003/Proceedings

R Core Team. (2016). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality*, *44*, 53–61. doi:10.1016/j.jrp.2009.10.005

Riefer, D. M. & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339. doi:10.1037/0033-295X.95.3.318

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184–201. doi:10.1037/1040-3590.14.2.184

Rost, J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324–332). Münster, Germany: Waxmann.

Rouder, J. N. & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604. doi:10.3758/BF03196750

Schmittmann, V. D., Dolan, C. V., Raijmakers, M. E. J., & Batchelder, W. H. (2010). Parameter identification in multinomial processing tree models. *Behavior Research Methods*, *42*, 836–846. doi:10.3758/BRM.42.3.836

Smith, J. B. & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. doi:10.1016/j.jmp.2009.06.007

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *64*, 583–639. doi:10.1111/1467-9868.00353

Thissen-Roe, A. & Thissen, D. (2013). A two-decision model for responses to Likert-type items. *Journal of Educational and Behavioral Statistics*, *38*, 522–547. doi:10.3102/1076998613481500

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*, 39–55. doi:10.1111/j.2044-8317.1990.tb00925.x

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*, 195–217. doi:10.1093/ijpor/eds021

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A steps model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). doi:10.1007/978-1-4757-2691-6_7

Weijters, B., Geuens, M., & Schillewaert, N. (2010a). The individual consistency of acquiescence and extreme response style in self-report questionnaires. *Applied Psychological Measurement*, *34*, 105–121. doi:10.1177/0146621609338593

Weijters, B., Geuens, M., & Schillewaert, N. (2010b). The stability of individual response styles. *Psychological Methods*, *15*, 96–110. doi:10.1037/a0018721

Wetzel, E., Carstensen, C. H., & Böhnke, J. R. (2013). Consistency of extreme response style and non-extreme response style across traits. *Journal of Research in Personality*, *47*, 178–189. doi:10.1016/j.jrp.2012.10.010

Zettler, I., Lang, J. W. B., Hülsheger, U. R., & Hilbig, B. E. (2016). Dissociating indifferent, directional, and extreme responding in personality data: Applying the three-process model to self- and observer reports. *Journal of Personality*, *84*, 461–472. doi:10.1111/jopy.12172

Appendix A

Recovery of ARS Parameters

In an additional simulation study, we checked that the decreased recovery of the ARS person parameters in Study 1 was due to the low absolute prevalence of ARS (and not due to any insufficiencies of the model or estimation method). For this purpose, we simulated data for two conditions that differed in the prevalence of acquiescence. The first condition with 20 items was identical to the first condition in Study 1 (see Figure 5). In the second condition, a higher prevalence of ARS was realized: The true ARS item parameters β_a were drawn from a distribution identical to that for MRS and ERS (i.e., β_m and β_e), that is, with $\mu_{\beta_a} = \Phi^{-1}(.70)$. The results illustrated in Figure A1 showed that the impaired recovery of the ARS-parameters in Figure 5 was mainly due to the low, but realistic prevalence of ARS and not due to insufficiencies of the Acquiescence Model. In contrast, the precision of the target-trait parameter decreased only slightly, indicating the Acquiescence Model provides unbiased estimates even when ARS is highly prevalent. Overall, this simulation shows the trade-off in measuring the target trait versus ARS responding—a trade-off that is theoretically predicted by the mixture structure illustrated in Figure 3.

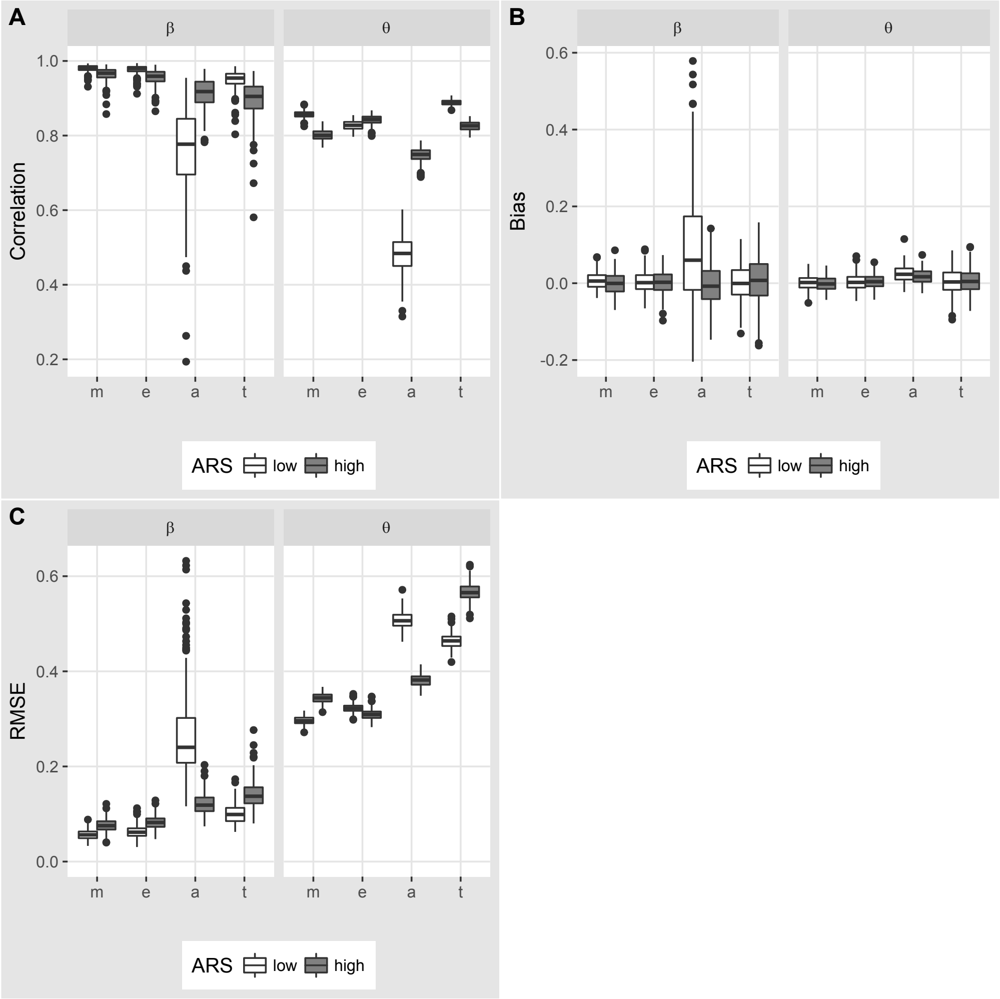


FIGURE A1: Boxplots illustrating parameter recovery of the item and person parameters of the Acquiescence Model when generating data with either a low (white) or high (gray) prevalence of acquiescence.

Appendix B

Graphical Convergence Diagnostic

When fitting the models both in the empirical study as well as in the simulation studies, careful attention was paid to convergence of the MCMC sampler. In this appendix, we focus on the model for all six HEXACO scales reported in the empirical study and showcase the convergence of four selected model parameters. For this illustration, we selected parameters that were of particular substantive interest and showed comparatively slow convergence. In particular, convergence is shown (a) for the lowest β_{aj}-parameter (i.e., the item most easily eliciting ARS responses, namely, $\beta_{a,67}$), (b) for the highest θ_{ai}-parameter (i.e., the person scoring highest on ARS, namely, $\theta_{a,294}$), (c) for the variance of the ARS person parameters σ_{θ_a} (i.e., $\mathbf{\Sigma}_{3,3}$), and (d) for the covariance between ARS and honesty-humility (i.e., $\mathbf{\Sigma}_{4,3}$). Note that both the θ- and the β-parameter are displayed on the original probit scale. For example, the estimate for $\beta_{a,67}$ on the probit scale is 1.09 $[0.89, 1.32]$, which corresponds to .86 [.81, .91] on the probability scale, which was reported above (see also Figure 6). Note further, that the estimate for $\mathbf{\Sigma}_{3,3}$ is also reported in Table 1, but the figure below shows the covariance $\mathbf{\Sigma}_{4,3}$ whereas Table 1 contains the corresponding correlation.

In all four panels in Figure B1, convergence is indicated by the following features: (a) the densities resulting from the six different chains are almost identical, (b) the "point estimates" as represented by the running mean are almost identical across chains, (c) the traceplots show nicely mixing chains without any irregularities, and (d) almost no autocorrelation is observed (a showcase of the capabilities of Stan).

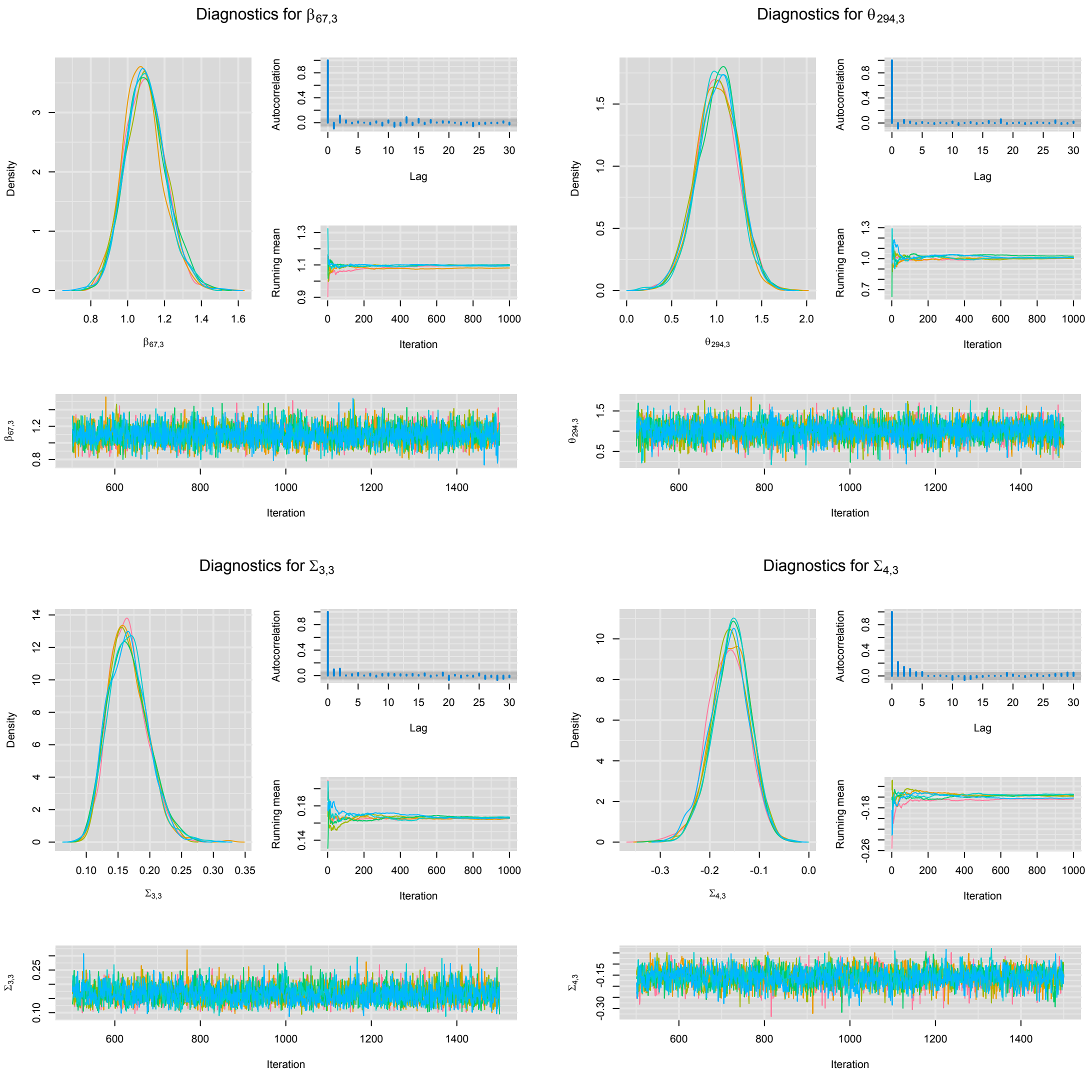


FIGURE B1: Graphical convergence diagnostic for four parameters. Displayed within each of the four panels are (in clockwise order starting from top left) a density plot, an autocorrelation plot, a plot of the running mean, and a traceplot (for each of six chains).

An Experimental Comparison of the Effect of Different Response Formats on Response Styles

Hansjörg Plieninger, Mirka Henninger, and Thorsten Meiser
University of Mannheim

Abstract

Many researchers fear that response styles threaten the quality of self-report data. Since post-hoc control of response style is sometimes difficult to implement in day-to-day usage, a-priori control by means of the response format is a promising alternative. In a recent experiment, Böckenholt (2017) found that response styles were less influential in a drag-and-drop format compared to a traditional Likert-type format. We aimed to study the replicability, causes, and generalizability of this effect and carried out two experiments. A similar effect was found in an almost exact replication of the original study. However, further conditions revealed that the effect was attributable to presenting the response categories in two columns rather than to the drag-and-drop format in general. In summary, we conclude that promoting a general advantage of a drag-and-drop format is premature. Rather, the Likert-type format remains unchallenged due to its easiness and robustness.

The data for the two conducted experiments are publicly available from https://osf.io/zu5th/.

Introduction

Response styles are defined as content-unrelated preferences for specific response categories. They have been discussed in the literature for decades, but still no consensus has been reached on how to deal with response styles in day-to-day data collection and analysis. The strategy that predominates in the psychometric literature these days is to correct post hoc for individual differences in response styles by means of a statistical model. However, a-priori control of response styles, for example, by means of the response format, may be effective as well and easier to accomplish in applied work. In a recent paper, Böckenholt (2017) proposed that a format akin to a card-sort task may be such a way of a-priori control. In the present study, we aimed to replicate this effect and investigate its generalizability and its underlying mechanism. In the remainder of the Introduction, we will discuss response styles and models thereof, different response formats, the original study by Böckenholt (2017), as well as the present research. Thereafter, we will present method and results for two experiments.

Response Styles

The most prominent response styles are the preference for extreme response categories (extreme response style, ERS), the preference for the middle category (midpoint response style, MRS), as well as the preference for affirmative response categories (acquiescence response style, ARS). These response styles have been discussed in numerous publications and many response style models and respective approaches have been proposed, especially in recent years. The models relevant to the present research will be briefly discussed in the following; for more comprehensive overviews see, for example, Böckenholt and Meiser (2017), Henninger and Meiser (2017), Wetzel, Böhnke, and Brown (2016), or Van Vaerenbergh and Thomas (2013). IR-tree models are multidimensional IRT models that assume a psychologically motivated tree-like structure of distinct processes in categorical data (Böckenholt, 2012; De Boeck & Partchev, 2012). For example, in order to capture MRS, the target trait, and ERS, a five-point item may be split into three binary decisions (i.e., pseudoitems): first, whether the midpoint was chosen or not; second, whether the respondent agreed or not; and third, whether an extreme response was given or not. IR-tree models for response styles have been validated (Plieninger & Meiser, 2014), extended (Jeon & De Boeck, 2016; Khorramdel & von Davier, 2014; Plieninger & Heck, 2017), and quickly became established in the methodological literature.

A different model class are multidimensional partial credit models (MPCMs) that have been adopted for response styles (Wetzel & Carstensen, 2017). Therein, the ordinal structure of a polytomous item is retained, but additional dimensions, for example, for ARS with specific coding schemes (e.g., [0 0 0 1 1]) are added to the model. A similar approach has been proposed in the framework of nominal response models (e.g., Bolt & Johnson, 2009; Bolt & Newton, 2011). IR-tree models and the MPCM can be used to both measure and control response styles but are faced with the problem that it is a challenge to disentangle response styles and content (especially when using few items).

So-called representative indicators of response styles (RIRS) are content-free measures of response styles (Couch & Keniston, 1960; De Beuckelaer, Weijters, & Rutten, 2010; Greenleaf, 1992). Therein, a large number of items each measuring a different construct is used. Given that the items are thematically unrelated, it is assumed that an individual response pattern (e.g., many extreme responses) cannot be attributed to the item content (alone), but is necessarily a measure of response style. Therefore, the items are recoded and aggregated (e.g., the extreme responses are counted) to form indicators of response styles (e.g., ERS). Those RIRS measures can be analyzed on their own (e.g., Weijters, Geuens, & Schillewaert, 2010), or the items can be analyzed jointly with a set of homogeneous items, for example, in an MPCM (Wetzel & Carstensen, 2017), in order to broaden the scope and enhance the reliability of the latent response style variables.

The current research—akin to Böckenholt (2017)—is based on two premises. First, interindividual differences in response styles exist and can be modeled. While this is common sense, it is nevertheless almost impossible to define normatively a state of "no" response styles. Thus, it is hard to judge whether, for example, more extreme responses are a sign of "more" response styles. However, more response styles in terms of more variability in response styles indicates that respondents use the response scale differently, and an increase in variability (but not in means) may lead to bias (e.g., Plieninger, 2016). Second, as a consequence, this (increased) response style variability may, if ignored, impair model fit and the measurement of the target trait and thus reliability and validity. Note, however, that these detrimental effects may be smaller than feared and sometimes even negligible (Plieninger, 2016; Savalei & Falk, 2014; Wetzel, Böhnke, & Rose, 2016).

Questionnaire Formats

Researchers and practitioners in psychology and beyond use multi-item questionnaires to measure a variety of attributes like self-reported personality or evaluation of a teacher, a subordinate, or a product. Such a questionnaire is comprised of the items, a response format, and a scoring scheme (which may not be disclosed).

Different response formats exists, and the Likert-type format is the most widely used one. This format offers, along with the item, a finite set of ordinal categories that represent gradual levels of agreement with the item content. The respondent has to choose one of the categories, which should be disjunct, exhaustive, and symmetric with respect to an explicit or implicit center. Oftentimes, between four and seven categories are used. This format is usually called Likert-type or rating format and is not to be confused with a genuine Likert scale[1].

The popularity of the Likert-type format is probably due to the fact that corresponding items are relatively easy to develop, administer, answer, score, and analyze. However, the analysis of corresponding data has also received a lot of controversial discussion. One issue is the scale level (ordinal vs. interval), which in turn determines the statistical models that may or may not be used. Another issue is that response styles may invalidate the data, which is the focus herein. Researchers have tried to address this, for instance, by modifying the number of categories, the use of appropriate category labels, the use of reverse-coded items, or the use of anchoring vignettes (e.g., von Davier, Shin, Khorramdel, & Stankov, 2017; Weijters, Cabooter, & Schillewaert, 2010; Weng, 2004). However, this may at best give partial control over response styles. Therefore, one may either choose to tolerate response styles assuming that their influence is negligible, or one may try to control response styles either post hoc by means of appropriate models or a priori. The response formats discussed in the following paragraphs may potentially be such a way of a-priori control, even though they have been developed and applied for numerous reasons.

The forced-choice format presents the respondent an k-tuple of $k \geq 2$ items, and the respondent is asked to rank them along some criterion, for example, “describes myself best to worst” (e.g., Brown & Maydeu-Olivares, 2012; Wetzel, Roberts, Fraley, & Brown, 2016). Thus, the format lacks any form of response categories, and respondents make relative judgments (compared to absolute judgments in the Likert-type format). Accordingly, it has been argued that the forced-choice format is less prone or even immune to response styles (Brown & Maydeu-Olivares, 2012). However, the relative judgments result in ipsative data that make it hard and sometimes impossible to make interindividual and not only intraindividual comparisons. Moreover, recent statistical models that aim to overcome the problems of ipsative data are difficult to apply in everyday use; and developing a forced-choice questionnaire is far more complex compared to the Likert-type format.

A closely related format is the so-called Q-Methodology (e.g., McKeown &

[1] A Likert scale is a measurement scale based on multiple items and constructed in a specific, extensive process. However, the Likert-type format is nowadays almost always used without the intent to develop a genuine Likert scale.

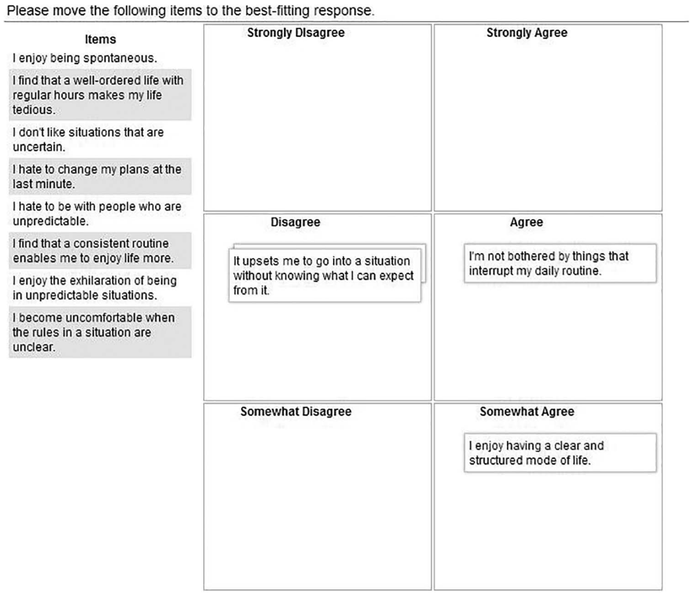
Thomas, 1988). Therein, respondents are presented all items at a time and are asked to sort them into a finite number of ordered categories. The set of categories usually mimics a normal distribution: For example, when respondents are asked to sort 25 items into a grid of 25 slots, the grid has one slot for category -4 and one for $+4$, two for -3 and two for $+ 3$, etc. Q-Methodology is said to foster comparative judgments and a holistic evaluation of the item pool. Furthermore, since the response distribution is fixed a priori, interindividual differences in category usage should have no effect. However, the data are ipsative in nature, and the current use of Q-Methodology does not offer interindividual comparisons.

In summary, the Likert-type format has the advantage that it produces absolute (not ipsative) data and that such questionnaires are easier to develop and analyze. The forced-choice format (as well as Q-Methodology) requires the comparison of multiple items, which is said (a) to facilitate a more holistic, thorough processing and (b) to be easier for respondents to make. A fourth format used by Böckenholt (2017) combines several advantages of these formats.

Drag-and-Drop Response Format

Böckenholt (2017) built on an idea of Thurstone (1928, 1930), who requested raters to rank a large number of items into 11 piles from "strongly in favor" of the topic to "strongly against" the topic. This method was used for scale construction, that is, for item selection and to calculate the scale value of each item. Note that in subsequent applications of the scale, an ordinary yes–no response format was employed. Böckenholt (2017) adapted Thurstone's format for computer-based administration and did not use it for scale construction but for genuine data collection. As illustrated in Figure 1, Böckenholt presented respondents a list of 12 items along with six response categories and asked participants to "move the ... items to the best-fitting response" via drag and drop (DnD).

In a between-participants experiment, Böckenholt (2017) compared the Likert-type, the DnD format, and a third format not discussed herein. In each condition, approximately 500 participants responded to 12 items measuring the construct personal need for structure (PNS; Neuberg & Newsom, 1993). Böckenholt fit two models to the data from each condition, namely, an ordinal graded response model (GRM; Samejima, 1969), which ignores response styles, as well as an IR-tree model, which accounts for the target trait, ERS, and MRS. According to AIC, the IR-tree model outperformed the GRM in the Likert condition, while the GRM fit better in the DnD condition. Böckenholt concluded: "The drag-and-drop method stands out because it triggered fewer response style effects than the other response formats. If this finding can be replicated in future research, one could argue that



Please move the following items to the best-fitting response.

Items

I enjoy being spontaneous.

I find that a well-ordered life with regular hours makes my life tedious.

I don't like situations that are uncertain.

I hate to change my plans at the last minute.

I hate to be with people who are unpredictable.

I find that a consistent routine enables me to enjoy life more.

I enjoy the exhilaration of being in unpredictable situations.

I become uncomfortable when the rules in a situation are unclear.

Strongly Disagree

Strongly Agree

Disagree

It upsets me to go into a situation without knowing what I can expect from it.

Agree

I'm not bothered by things that interrupt my daily routine.

Somewhat Disagree

Somewhat Agree

I enjoy having a clear and structured mode of life.

FIGURE 1: Drag-and-drop response format used by Böckenholt (2017). From "Measuring response styles in Likert items," by Ulf Böckenholt, 2017, *Psychological Methods, 22*, p. 75. Copyright (2016) by the American Psychological Association. Reprinted with permission.

had Thurstone's (1928) approach been adopted instead of Likert's (1932) approach, response styles would play a much smaller role than they do now" (2017, p. 80).

Böckenholt's (2017) findings are promising for several reasons. First, the drag-and-drop format is a method to control response style a priori. This means that it may no longer be necessary to correct the data post hoc. This would be advantageous for low-stakes applications, because guidelines for selecting one of the many available methods and applying it do hardly exist. Second, DnD data can be dealt with just like Likert-type data because both are absolute (not relative) judgments. Thus, standard statistical procedures tried and tested with Likert-type data can be applied to DnD data. Third, the method is relatively easy to apply in both computer- and paper-based settings.

However, several questions remain for further research. First, the psychological mechanism that led to the observed results is unknown. Böckenholt offers two possible explanations: "The novelty effect could have increased attention to the item content and hence elicited more careful responses. Moreover, the comparisons with other items in the same response category may have facilitated comparative processes that may not be triggered when comparing an item with a response category alone, further reducing the incidence of response-style effects" (2017, p. 80). Second, the relative model fit of an IR-tree model and an ordinal GRM revealed an advantage of the drag-and-drop format. It would be interesting to examine this effect with respect to other models and to potential consequences such as reliability and validity. Third, it is important to assess the generalizability of the effect with respect to other items and constructs, other response scales (i.e., number of categories, anchors), other populations, other response styles (especially ARS), and other measures of and approaches to response styles. The issue of generalizability receives additional relevance since the data in the Likert condition of Böckenholt (2017) showed a specific response distribution (see also Appendix B).

The Present Research

The aim of the present research was to replicate the findings of Böckenholt (2017), to study the generalizability of the described effect, and to investigate the two proposed psychological mechanisms. In Study 1, we aimed to contrast the mechanisms of novelty and comparative processes and designed our experiment accordingly. Three response formats were examined: In the Likert condition, responses were given using a standard format probably familiar to many respondents. In the second and third condition, responses were given using drag and drop similar to the format used by Böckenholt. However, in the condition DnD-open (see Figure 2A), respondents could see all previously answered items below each other (in the response category they selected), which should have facilitated comparative processes. In the condition DnD-shut in contrast (see Figure 2B), items already answered were masked, which should have hindered comparative processes. Taken together (see Table 1), the novelty explanation predicts that response styles should be less influential in the two DnD conditions compared to the Likert condition, because both DnD formats—in contrast to the Likert-type format—are probably unfamiliar to respondents. The comparison explanation predicts that response styles should be less influential in the DnD-open condition compared to both the Likert and the DnD-shut condition, because the DnD-open condition allows the respondent to easily compare the current item to previously answered items. Such comparisons are more demanding in the Likert condition and very cumbersome in the DnD-shut condition. In the two DnD conditions, response categories were presented in only one column in order to emphasize the gradual ordering of the categories and because five instead of six categories were used.

In Study 2, we focused on the fact that the response categories in the study by Böckenholt (2017) were arranged in two columns with agreement on the right and disagreement on the left (see Figure 1). We implemented this format in the condition DnD-II and contrasted it with a format with a single column of response categories, DnD-I (see below). In this latter format, the gradual ordering from disagreement to agreement is more explicit. In both conditions, the last item dropped in each category was permanently visible, and this should have made comparative processes as easy as in the original study and easier than in the Likert and the DnD-shut conditions. If the mechanism underlying the original effect is related to novelty or comparisons, the effect should replicate in both DnD-I and DnD-II. However, if the mechanism (whatever it may be) is related to the 2-column format rather than drag and drop, it should replicate only in the DnD-II condition (see Table 1).

Table 1: Summary of Hypotheses

	Condition				
	Study 1		Study 2		1+2
Mechanism	DnD-open	DnD-shut	DnD-I	DnD-II	Likert
Novelty Effect	−	−	−	−	+
Comparative Processes	−	+	−	−	+
Related to 2-column format	+	+	+	−	+

Note. + and − indicate high and low influence of response styles as predicted by the respective mechanism in the respective condition.

In both experiments, the aim was to examine the effect of response format from multiple perspectives. In addition to the comparison of an IR-tree model and a GRM replicating the original analysis, we pursued the following routes. First, the RIRS method was used to obtain additional, content-free measures of response styles. Second, an MPCM was used as an alternative model to the IR-tree model, because it retains the ordinal character of the response scale. Third, RIRS were added to the MPCM to obtain more reliable, model-based estimates of response styles. Fourth, the reliability and validity of two scales was compared across conditions. Overall, if the original finding generalized to the conducted experiments, response styles should be less influential in the DnD conditions in all analyses. This could, for example, mean that the ERS variance in an MPCM should be largest in the Likert condition, or that reliability and validity should be largest when using drag and drop.

Study 1

Method

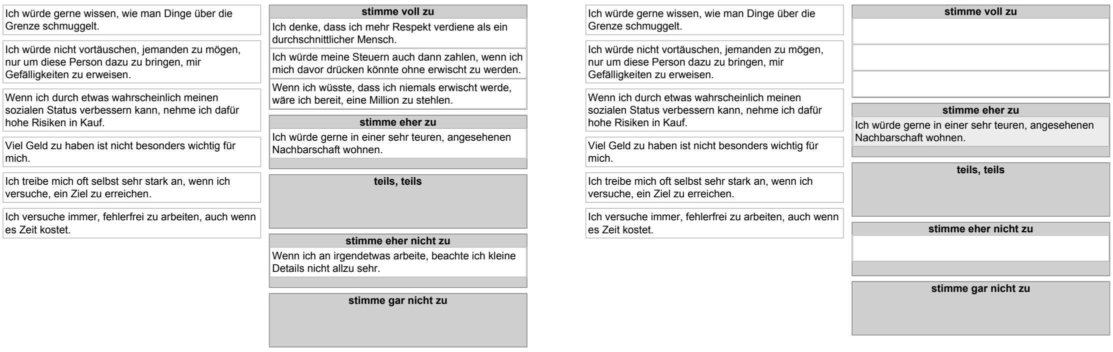
Participants

A convenience sample of German participants was recruited both online ($n = 551$) and in the laboratory ($n = 93$) in April and May 2017 with an analysis sample size of $n = 644$. Participants were on average 28 years old ($SD = 10$, range from 18 to 80) and 74 % of them were female. Online, participants could win one of

10 vouchers worth of 20 €, while each of the participants in the laboratory was compensated with either 2 € or course credit. Seventy-three participants were excluded beforehand who failed at least one of two instructional manipulation checks (Meade & Craig, 2012; Oppenheimer, Meyvis, & Davidenko, 2009), or who took on average less than two seconds per item (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Meade & Craig, 2012), or who indicated that their data should not be used (Meade & Craig, 2012).

Materials and Procedure

Both in the lab and online, participants were randomly assigned to one of three conditions, namely, Likert (n = 203), DnD-open (n = 218), and DnD-shut (n = 223). In the Likert condition, participants answered the items using a standard rating format almost identical to the one used by Böckenholt (2017). In the DnD conditions (see Figure 2), participants responded to the items by dragging the item from the list of items into the desired category. Previously answered items were stacked underneath each other (in the respective category), and the response categories increased in vertical size to enclose all respective items. In the DnD-open condition, all items were permanently visible, and we assumed that this would facilitate comparisons among items. In the DnD-shut condition, already answered items were masked and could be made temporarily visible upon mouseover.



Ich würde gerne wissen, wie man Dinge über die Grenze schmuggelt.

Ich würde nicht vortäuschen, jemanden zu mögen, nur um diese Person dazu zu bringen, mir Gefälligkeiten zu erweisen.

Wenn ich durch etwas wahrscheinlich meinen sozialen Status verbessern kann, nehme ich dafür hohe Risiken in Kauf.

Viel Geld zu haben ist nicht besonders wichtig für mich.

Ich treibe mich oft selbst sehr stark an, wenn ich versuche, ein Ziel zu erreichen.

Ich versuche immer, fehlerfrei zu arbeiten, auch wenn es Zeit kostet.

stimme voll zu

Ich denke, dass ich mehr Respekt verdiene als ein durchschnittlicher Mensch.

Ich würde meine Steuern auch dann zahlen, wenn ich mich davor drücken könnte ohne erwischt zu werden.

Wenn ich wüsste, dass ich niemals erwischt werde, wäre ich bereit, eine Million zu stehlen.

stimme eher zu

Ich würde gerne in einer sehr teuren, angesehenen Nachbarschaft wohnen.

teils, teils

stimme eher nicht zu

Wenn ich an irgendetwas arbeite, beachte ich kleine Details nicht allzu sehr.

stimme gar nicht zu

(A) Response format DnD-open.

Ich würde gerne wissen, wie man Dinge über die Grenze schmuggelt.

Ich würde nicht vortäuschen, jemanden zu mögen, nur um diese Person dazu zu bringen, mir Gefälligkeiten zu erweisen.

Wenn ich durch etwas wahrscheinlich meinen sozialen Status verbessern kann, nehme ich dafür hohe Risiken in Kauf.

Viel Geld zu haben ist nicht besonders wichtig für mich.

Ich treibe mich oft selbst sehr stark an, wenn ich versuche, ein Ziel zu erreichen.

Ich versuche immer, fehlerfrei zu arbeiten, auch wenn es Zeit kostet.

stimme voll zu

stimme eher zu

Ich würde gerne in einer sehr teuren, angesehenen Nachbarschaft wohnen.

teils, teils

stimme eher nicht zu

stimme gar nicht zu

(B) Response format DnD-shut.

Figure 2: Drag-and-drop formats used in Study 1. To-be answered items appear on the left, and already answered items appear in the chosen category on the right.

Respondents provided informed consent and then answered two test items in the assigned conditions to familiarize themselves with the response format. Then, participants responded to 42 personality items from the HEXACO inventory (four pages), 41 heterogeneous items (four pages), a behavioral measure not affected by response styles, namely, a dictator game, and finally demographic questions.

Measures

All items were administered using five response categories labeled “strongly agree”, “rather agree”, “yes and no”, “rather disagree”, and “strongly disagree”.

HEXACO We used 42 items from the 200-item HEXACO inventory (Lee & Ashton, 2006), namely, its German translation; 24 items were reverse-coded. In detail, we measured two conscientiousness facets, namely, prudence (α = .79) and diligence (α = .82) as well as two honesty-humility facets, namely, fairness (α = .76) and greed avoidance (α = .77) with eight items each (i.e., 32 items in total). Additionally, we assessed the factors conscientiousness (α = .80) and honesty-humility (α = .71) with 10 items each (as in the standard 60-item HEXACO inventory); since there is overlap between the facets and the factors, only 10 additional items had to be included. That means, either the 32 facet items or the 20 factor items may be used in the analysis but not all 42 items at once. The items were selected to predict external criteria (see below) and to use them in psychometric models. The items were randomly assigned to one of four pages (balancing domain and coding direction), but the order was the same for all participants.

Heterogeneous items In order to obtain content-free measures of response styles, 41 items were selected from various scientific questionnaires. From each multi-item scale, one item was selected: either an item of intermediate difficulty or, if such information was not available, at random. Post hoc, two items were removed that showed absolute bivariate correlations $>$ .40. The remaining 39 items had correlations ranging from $-$.36 to .40 with an average correlation of .04 and an average absolute correlation of .09. Item difficulties (possible range from 0 to 1, i.e., easy to hard) ranged from .31 to .77 with an average of .53 (SD = .11). A coding scheme of (0, 0, 1, 0, 0) was used to form indicators for MRS, (1, 1/2, 0, 1/2, 1) for ERS, and (0, 0, 0, 1, 1) for ARS. Subsequently, the mean across the respective indicators was used as a content-free measure of response style. On average, MRS responses were given with a prevalence of .27 (SD = .10), ERS had a mean of .48 (SD = .11), and ARS had a mean of .39 (SD = .11).

Dictator Game In a *hypothetical* dictator game (e.g., Hilbig, Thielmann, Hepp, Klein, & Zettler, 2015), respondents were informed that they were provided with 10 € and that they could freely decide how much of that money they wanted to give to an unknown other. The allocated money is typically interpreted as a measure of prosociality, and it is empirically related to honesty-humility with a meta-analytic estimate of r = .24 (Zhao & Smillie, 2015; Hilbig et al., 2015). Based on the premise that behavioral measures such as dictator-game offers were independent of response styles and of the experimental manipulation, we investigated the relationship with honesty-humility across conditions. If the new response format indeed reduced the influence of response styles and thereby improved the measurement of honesty-humility, this should increase validity and thus result in a higher correlation with dictator-game offers.

Demographics Respondents provided their age, sex, years of education, height, and weight, and their body mass index (BMI) was calculated. Research consistently revealed a small, negative correlation between BMI and conscientiousness (or facets thereof); higher levels of conscientiousness are presumably accompanied with dietary and exercise habits that protect against high BMI levels, which may explain this finding (e.g., Brummett et al., 2006; Sutin, Ferrucci, Zonderman, & Terracciano, 2011). Again, it was assumed that BMI was a response-style-free measure, and we tested whether the new response format improved the validity, namely, resulted in a stronger, more negative, relationship between BMI and conscientiousness.

Results

As outlined above, we aimed to look at the influence of response formats on response styles from multiple perspectives. Therefore, different routes were pursued and the respective results will be reported in the following. IR-tree models were fit using Mplus 7.4 (Muthén & Muthén, 2012), and all other analyses were carried out in R 3.4.2 (R Core Team, 2017)[2]. In the analyses that directly compared the three groups, we contrasted each of the two DnD conditions with the Likert condition via dummy coding.[3] If applicable, a significance level of .05 was used.

Descriptive statistics of the observed variables are reported in Table A1 of the Appendix. The DnD formats were a little bit more time consuming with median times across the eight questionnaire pages of 8.5 minutes (Likert), 9.5 (DnD-open), and 9.1 (DnD-shut), respectively.

IR-Tree Models

Herein, we focused on the two factors honesty-humility and conscientiousness and specified the models analogously to Böckenholt (2017)[4]. As a baseline model

[2] We made use of the R packages psych (Revelle, 2017), TAM (Robitzsch, Kiefer, & Wu, 2017), and tidyverse (Wickham, 2017).

[3] Theoretically, contrast coding would better reflect our hypotheses in Table 1, but—running the analyses—we realized that our results are more meaningfully described using dummy coding.

[4] That is, the discrimination parameter pertaining to the target trait(s) were freely estimated and those pertaining to response styles were fixed to 1. A probit link was used. And both MRS

without response styles, a GRM with two correlated dimensions was fit to the 20 items. This model was contrasted with an IR-tree model. Note that we used only five (instead of six) response categories, and therefore a different IR-tree model (Böckenholt, 2012) was used, which was described in the Introduction above. This model captured MRS, ERS, and the two target traits. To facilitate interpretation, we report AIC ratios (i.e., $\text{AIC}_{\text{IR-tree}}/\text{AIC}_{\text{GRM}}$), where a ratio < 1 favors the IR-tree model over the GRM, and a ratio > 1 favors the GRM. As shown in Figure 3, Böckenholt (2017) found that an IR-tree model outperformed a GRM in the Likert condition (i.e., ratio < 1) but not in the DnD condition (i.e., ratio > 1), where response styles were less influential. In our data, however, the opposite pattern was found. The GRM was most favored in the Likert condition, which seemed to be least affected by response styles. A similar pattern was found in the DnD-shut condition, whereas the effect was reversed in the DnD-open condition, where the IR-tree model was favored.

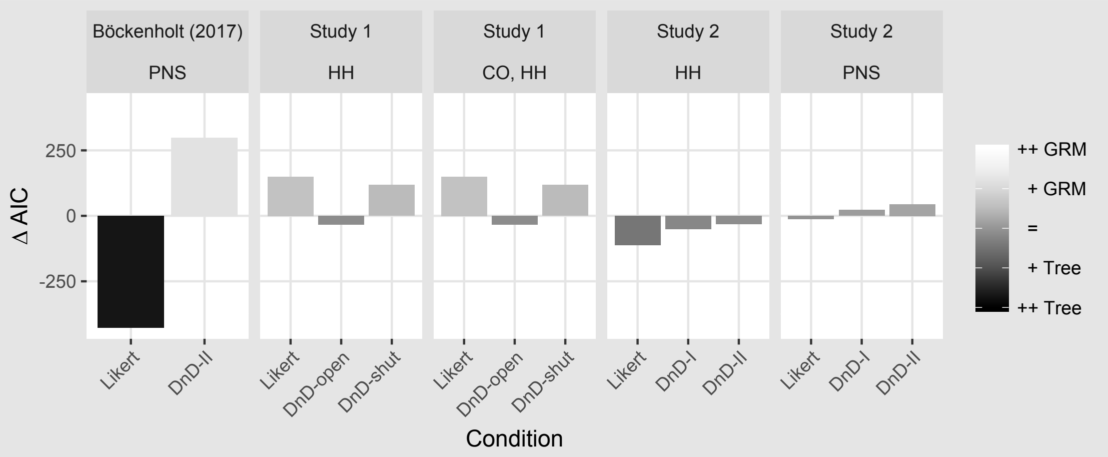


FIGURE 3: Relative model fit in terms of AIC of an IR-tree model and a GRM across conditions in Study 1 and 2 and in Böckenholt (2017). Ratios > 1 (< 1) appear in light (dark) gray and favor a GRM without response styles (an IR-tree model with response styles).

We performed a number of robustness analyses: Different items were used, namely, either only the honesty-humility items (see Figure 3), or only the conscientiousness items, or the 32 facet items (and the appropriate number of target traits was specified); BIC instead of AIC was used as a criterion; all item discrimination parameters were fixed to 1. None of these analyses changed our interpretation, because the GRM was almost always most favored in the Likert condition, which is in sharp contrast to the findings of Böckenholt (2017).

and ERS were assumed to be unidimensional, whereas two dimensions were assumed content-wise, one for each underlying trait.

Heterogeneous Items

To obtain an alternative, content-free measure of response styles, the heterogeneous items were recoded and aggregated as described above. When comparing ERS across condition, we found that the ERS level was lowest in the Likert condition with an average of 0.47. A linear model revealed that more ERS responses were given in the DnD-shut condition ($b = 0.03, p = .005, d' = 0.27$), but not in the DnD-open condition ($p = .156$). For ARS and MRS, no significant mean differences were found, and no significant difference in variances was found for any of the three measures using a Brown–Forsythe test. Note again that it is difficult to define a normatively optimal level of these measures. For example, is a level of 20 % or 30 % of midpoint responses more "appropriate"? Nevertheless, we conclude that all three response formats elicited response styles to similar degrees and that the DnD-open format led to slightly more ERS responses.

Multidimensional Partial Credit Model

We fit a single MPCM to the data from all three conditions and included the 20 HEXACO items as well as the 39 heterogeneous items. Four latent variables were specified, one measuring honesty-humility (10 items), one measuring conscientiousness (10 items), one for ERS, and one for ARS (each measured by all 59 items). Note that we deliberately did not include MRS since it is highly collinear to ERS here. In a multi-group variant of the model, the means and (co-)variances of these latent variables were allowed to vary across conditions.[5] The results, displayed in Table 2, revealed only small descriptive differences between response formats: Extreme responses were given less often in the DnD-open condition, and a little bit more ARS variance was observed in the DnD-shut condition. These results extend those from the previous section (on heterogeneous items), because 20 additional items were used (i.e., more information to obtain precise estimates), and because an IRT model was employed. Furthermore, the scope of ARS is extended: Being high on ARS now also relates to agreement with both regular and reverse-coded items (while reverse coding is a meaningless concept in case of heterogeneous items).

The same pattern as in Table 2 was observed when using the 32 facet items (for the four facets) instead of the 20 factor items (for the two factors). In summary, this led to the conclusion that the effect of ERS was comparable across conditions while ARS was somewhat more influential in the DnD-shut condition.

[5] Unfortunately, respective standard errors are not available when freeing the (co-)variances using the function `tam.mml.3pl()` (Robitzsch et al., 2017).

Table 2: Effect of Response Formats on Latent Response Style Means and Variances

		Means		Variances	
Study	Condition	ERS	ARS	ERS	ARS
Study 1	Likert	0.00[a]	0.00[a]	0.19	0.14
	DnD-open	-0.08	-0.06	0.18	0.13
	DnD-shut	0.04	-0.07	0.16	0.20
Study 2	Likert	0.00[a]	0.00[a]	0.27	0.08
	DnD-I	-0.03	-0.05	0.37	0.09
	DnD-II	0.18	-0.04	0.25	0.08

[a] Set to zero for identification purposes (i.e., dummy coding).

Reliability and Validity

We used Cronbach's alpha as an indicator of the reliability of the HEXACO scales and compared it across conditions (see Table 3). Each of the two DnD conditions was tested against the Likert condition for significance (Duhachek & Iacobucci, 2004). Note that the results were not corrected for multiple comparisons and that the factors and the facets are not independent (because they share some items); therefore, a single finding should be treated with caution. Nevertheless, the alphas in the Likert condition seemed at least as high as in the two DnD conditions if not higher, while the two DnD conditions led to similar results. This advantage of the Likert condition contradicted our expectations.

Table 3: Influence of Response Formats on Cronbach's Alphas of the HEXACO Scales

		Cronbach's Alpha		
Scale	Items	Likert	DnD-open	DnD-shut
Honesty-humility	10	.69	.70	.73
Fairness	8	.77	.74	.77
Greed avoidance	8	.79	.79	.74$^+$
Conscientiousness	10	.83	.77*	.79$^+$
Diligence	8	.86	.79**	.81$^+$
Prudence	8	.79	.79	.79

Note. Asterisks indicate results of pairwise comparisons (two-sided) between the Likert condition and a drag-and-drop (DnD) condition. $^+p < .10$. $^*p < .05$. $^{**}p < .01$.

The correlations of the HEXACO scales with external criteria served as indicators of validity and were compared across conditions. Research showed that honesty-humility (and/or its facets) is positively related to offers in a dictator game, and that conscientiousness (and/or its facets) is negatively related to BMI. Thus, an increase in “true” variance and a decrease in response style variance should lead to stronger correlations. The correlation between honesty-humility and dictator-game offer was significantly positive in all three conditions (see Table 4). The magnitude of the relationship, however, did not differ across condition (i.e., non-significant interactions in a linear model). The correlation between conscientiousness and BMI offer was significantly negative (one-tailed) only in the Likert condition, whereas it was non-significant in the two DnD conditions. However, the difference between conditions was non-significant (i.e., no interactions in a linear model).

Table 4: Influence of Response Formats on Validity Correlations

			Correlation	
Study	Condition		DG	BMI
Study 1	Likert	HH	.26	
	DnD-open	HH	.27	
	DnD-shut	HH	.32	
	Likert	CO		−.13
	DnD-open	CO		−.01
	DnD-shut	CO		.01
Study 2	Likert	HH	.23	
	DnD-I	HH	.14	
	DnD-II	HH	.29	

Note. DG = dictator-game offer; BMI = body mass index; CO = conscientiousness; HH = honesty-humility.

We conducted several robustness analyses and (a) replaced the HEXACO factors by the respective facets, (b) added additional main effects of sex, age, and education to the linear models, and/or (c) log-transformed BMI. None of the analyses changed our interpretation, but the relationship between BMI and the HEXACO facets (diligence or prudence) was even more in favor of the Likert condition. In summary, both predicted validity coefficients were observed in the Likert condition, while only one effect was found in the two DnD conditions. This led to the unexpected conclusion that the Likert condition worked best.

Summary of Study 1

We did not find—in contrast to the study of Böckenholt (2017)—that a DnD format was less affected by response styles than a Likert-type format. Even though the effects were not very large, the Likert-type format outperformed the two DnD formats consistently across a wide range of analyses. In order to explain differences between our results and those reported by Böckenholt, we conducted another experiment, which served two purposes: First, a more direct replication was carried out in that—in contrast to Study 1—six response categories were offered and the PNS items were used. Second, we investigated a further, potential explanation for the advantage of DnD over Likert, namely, the fact that the response categories were presented in two columns in the study of Böckenholt as shown in Figure 1 and not in one column (see Figure 2). We had no clear hypothesis, but the more compact display with two columns, the clearer distinction between agreement and disagreement, and the less obvious gradual ordering of categories may have led to response processes that had an effect on the considered outcomes.

Study 2

Method

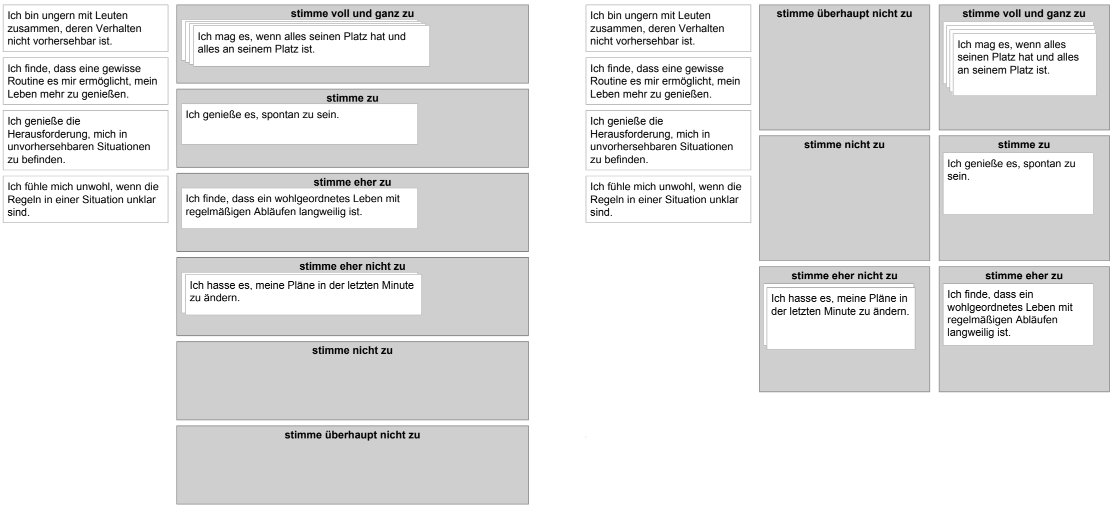
Participants

A sample of German participants was recruited by means of the non-representative online panel SoSci Survey in October 2017. The analysis sample was comprised of $n = 506$ participants who were on average 39 years old ($SD = 15$, range from 18 to 83) and 58 % of them were female. Participants could win one of two vouchers worth of 50 €. One-hundred and six participants were excluded beforehand who failed at least one of two instructional manipulation checks (Meade & Craig, 2012; Oppenheimer et al., 2009), or who took on average less than two seconds per item (Huang et al., 2012; Meade & Craig, 2012), or who indicated that their data should not be used (Meade & Craig, 2012).

Materials and Procedure

Participants were randomly assigned to one of three conditions, namely, Likert ($n = 181$), drag-and-drop one column ("DnD-I"; $n = 162$), and drag-and-drop two column ("DnD-II"; $n = 163$). The Likert condition was almost identical to the one used by Böckenholt (2017) and the one in Study 1. In the two DnD conditions (see Figure 4), participants indicated their response by dragging the item from the list of items and dropping it into the selected category. The six categories were either presented in a single column ("DnD-I") or in two columns ("DnD-II") as in the condition of Böckenholt shown in Figure 1.

Respondents provided informed consent and then answered two test items in order to familiarize themselves with the response format. Then, participants re-



Ich bin ungern mit Leuten zusammen, deren Verhalten nicht vorhersehbar ist.

Ich finde, dass eine gewisse Routine es mir ermöglicht, mein Leben mehr zu genießen.

Ich genieße die Herausforderung, mich in unvorhersehbaren Situationen zu befinden.

Ich fühle mich unwohl, wenn die Regeln in einer Situation unklar sind.

Ich mag es, wenn alles seinen Platz hat und alles an seinem Platz ist.

Ich genieße es, spontan zu sein.

Ich finde, dass ein wohlgeordnetes Leben mit regelmäßigen Abläufen langweilig ist.

Ich hasse es, meine Pläne in der letzten Minute zu ändern.

(A) Response format DnD-I.

Ich bin ungern mit Leuten zusammen, deren Verhalten nicht vorhersehbar ist.

Ich finde, dass eine gewisse Routine es mir ermöglicht, mein Leben mehr zu genießen.

Ich genieße die Herausforderung, mich in unvorhersehbaren Situationen zu befinden.

Ich fühle mich unwohl, wenn die Regeln in einer Situation unklar sind.

Ich mag es, wenn alles seinen Platz hat und alles an seinem Platz ist.

Ich genieße es, spontan zu sein.

Ich hasse es, meine Pläne in der letzten Minute zu ändern.

Ich finde, dass ein wohlgeordnetes Leben mit regelmäßigen Abläufen langweilig ist.

(B) Response format DnD-II.

FIGURE 4: Drag-and-drop formats used in Study 2. To-be answered items appear on the left, and already answered items appear in the chosen category on the right.

sponded to 12 PNS items (one page), 10 HEXACO items (one page), 32 heterogeneous items (three pages), the dictator game, and demographic questions.

Measures

All items were administered using six response categories labeled "strongly agree", "agree", "rather agree", "rather disagree", "disagree", and "strongly disagree".

PNS Parallel to Böckenholt (2017), respondents answered the 12 PNS items (Neuberg & Newsom, 1993), namely the German translation (Machunsky & Meiser, 2006), and only 11 items were used in the analyses. These items can be used to measure a global scale ($\alpha = .80$), or to measure the facets desire for structure ($\alpha = .72$) and response to lack of structure ($\alpha = .75$).

HEXACO Similar to Study 1, we administered the 10 items measuring honesty-humility ($\alpha = .70$; Lee & Ashton, 2006; Moshagen, Hilbig, & Zettler, 2014).

Heterogeneous items Again, heterogeneous items were used, but only 32 due to external restrictions. Post hoc, one item was removed that showed an absolute, bivariate correlation $> .40$. The remaining 31 items had correlations ranging from $-.37$ to $.34$ with an average correlation of .03 and an average absolute correlation of .09. Item difficulties ranged from .33 to .75 with an average of .51 ($SD = .11$). A coding scheme of $(0, 0, 1, 1, 0, 0)$ was used to form indicators for MRS, $(1, 1/2, 0, 0, 1/2, 1)$ for ERS, and $(0, 0, 0, 1, 1, 1)$ for ARS. Subsequently, the mean across the respective indicators was used as a content-free measure of response style. On average, MRS responses were given with a prevalence of .47 ($SD = .17$), ERS had a mean of .35 ($SD = .14$), and ARS had a mean of .54 ($SD = .11$).

Dictator Game The employed dictator game was identical to that used in Study 1 described above.

Results

Descriptive statistics of the observed variables are reported in Table A2 of the Appendix. There were no mean differences of honesty-humility and PNS across conditions, but honesty-humility showed more variability in DnD-II compared to the Likert condition ($p = .042$)—a pattern that held descriptively also in comparison to DnD-I and for PNS. This effect was not only found between persons, but also within persons[6] meaning that respondents used a wider range of categories with the DnD-II format. Interestingly, respondents took about equally long for the five questionnaire pages with medians of 7.6 (Likert) and 7.7 minutes (DnD-I and DnD-II).

IR-Tree Models

First, an IR-tree model and a GRM were fit to the PNS data from each condition as a direct replication of the analyses of Böckenholt (2017). In comparing the relative fit of these two models across conditions, we were able to replicate the pattern found by Böckenholt: A GRM outperformed an IR-tree model in the DnD-II condition with a reversed pattern in the Likert condition (see Figure 3). However, the differences between models were smaller compared to the original study (and according to BIC, the GRM was preferred in all three conditions). Moreover, the advantage of the GRM was less pronounced in the DnD-I condition. Second, the same model was fit to the honesty-humility items, and there—even though the relative differences between models were similar—the IR-tree model was preferred in all three conditions (see Figure 3). Thus it seemed that the effect reported by Böckenholt did not generalize to the honesty-humility items used herein.

Heterogeneous Items

To obtain an alternative, content-free measure of response styles, the heterogeneous items were recoded and aggregated as described above. When comparing ERS across condition, we found that the ERS level was lowest in the Likert condition with a mean of 0.34. A linear model revealed no difference to the

[6] For PNS, the within-person SD averaged 1.13 (Likert) compared to 1.14 (DnD-I, $p = .740$), and 1.28 (DnD-II, $p < .001, d' = 0.43$)).

DnD-I condition ($p = .859$), but more ERS responses in the DnD-II condition ($b = 0.04, p = .010, d' = 0.28$). Like a mirror image, MRS was lower in the DnD-II condition ($b = -0.04, p = .013, d' = -0.27$) compared to Likert but not DnD-I. The number of ARS responses did not differ across conditions, and there were no significant differences in ERS or ARS variance across conditions as revealed by a Brown–Forsythe test.

Multidimensional Partial Credit Model

We fit a single MPCM to the data from all three conditions and included the PNS and honesty-humility items as well as the 31 heterogeneous items. Five latent variables were specified, namely, desire for structure (4 items), response to lack of structure (7 items), honesty-humility (10 items), as well as ERS, and ARS (each measured by all 52 items). Again, MRS was not included because of its collinearity to ERS. In a multi-group variant of the model, the means and (co-)variances of the latent variables were allowed to vary across conditions.

The results, displayed in Table 2, revealed that more ERS responses were given in the DnD-II condition, and that the ARS level was roughly the same across conditions. Interestingly, ERS variance was smallest in the DnD-II condition (which can not be attributed to a ceiling effect) and largest in the DnD-I condition.[7] Thus, we concluded that the increase in variability of observed scores for DnD-II reported above (see also Table A2) seemed to be attributable to "true" variability instead of merely stylistic variability.

Reliability and Validity

The alphas in the three conditions were highly similar for PNS with $\alpha = .81$ (Likert), $\alpha = .77$ (DnD-I), and $\alpha = .81$ (DnD-II) as well as for honesty-humility with with $\alpha = .67$ (Likert), $\alpha = .68$ (DnD-I), and $\alpha = .73$ (DnD-II). The pairwise comparisons between the Likert condition and each respective DnD condition were non-significant for both scales.

The correlation between honesty-humility and dictator-game offer was significantly (one-tailed) positive in all three conditions (see Table 4). The magnitude of the relationship, however, did not differ across condition (i.e., non-significant interaction in a linear model) even though the effect in the DnD-II condition was twice as large as in the DnD-I condition.

In short, the reliability and validity did not differ significantly across conditions even though there was a small, descriptive advantage of the DnD-II condition.

[7] This holds also relative to the variability of all five latent variables (i.e., $\sigma^2_{\text{RS}}/tr(\mathbf{\Sigma})$).

Summary of Study 2

We were able to replicate the finding that response styles are less influential in a DnD-II format compared to a Likert-type format. In a variety of analyses, this advantage was found consistently even though it was rather small. However, this advantage was not found in a DnD-I format, which appeared to be comparable if not inferior to a Likert-type format in terms of response style effects, reliability, and validity.

Discussion

The aim of the present work was to investigate the effect of an innovative response format on response styles. A recent study by Böckenholt (2017) revealed that such a format may offer—in contrast to a Likert-type format—better control over response styles. This effect was promising, because statistical control of response styles post hoc, after data collection, is a complex, controversial, and unresolved problem. In contrast, controlling response styles a priori by means of the questionnaire and response format is something that other researchers have tried to address even though with limited success for day-to-day usage. Therefore, we aimed to replicate the effect of response formats, to investigate the underlying mechanism, and to examine its generalizability with the hope to help establish a new response format. This format requires respondents to manually sort a list of items into a list of categories (via drag and drop). Böckenholt conjectured that the effect on response styles may be either due to the novelty of the format or to comparisons among items that are easier to perform in the new format.

Study 1 was a between-participants experiment to test the two proposed mechanisms. According to the novelty effect, response styles should be less influential in both DnD conditions compared to the Likert condition. In contrast, comparative processes should lead to an advantage of the DnD-open condition compared to both DnD-shut and Likert. In short, the results were in sharp contrast to our expectations, because response styles were least influential in the Likert condition. This was revealed consistently across a wide range of analyses: A graded response model (without response styles) was most favored over an IR-tree model in the Likert condition. Furthermore, ARS variance was highest in the DnD-shut condition as revealed by an alternative IRT model and content-free measures of response styles. Moreover, the reliability and validity of the investigated measures was as least as high if not higher in the Likert condition. While each of these analyses alone may have its limitation, the results in summary consistently revealed that the Likert format performed as least as good if not better in comparison to the DnD formats. The differences between the DnD-open and the DnD-shut conditions were small and inconsistent.

The failure to replicate the effect reported by Böckenholt (2017) may possibly be attributable to sample characteristics, the items used (HEXACO vs. PNS), the number of categories (5 vs. 6), or subtle differences in the implemented drag-and-drop format. We therefore designed Study 2 to rule out these explanations (except sample characteristics), and to test a third potential explanation, namely, whether the effect may be attributable to presenting the categories in two columns. Again, a between-participants experiment was carried out and our condition DnD-II mirrored the original condition as closely as possible. In a second condition, DnD-I, categories were presented in only one column parallel to the DnD formats of Study 1 and more similar to Likert-type formats. We were able to replicate the original effect of an advantage of DnD-II over Likert. Even though this advantage was consistently found, it was not large. For example, reliabilities and validities were comparable, ERS and ARS variance in an MPCM were comparable, and the advantage of a GRM over an IR-tree model was only found for PNS items and not for honesty-humility items. Apart from that, the effect vanished almost completely in the DnD-I format, which had lowest reliability and validity and highest ERS variance (in an MPCM). These findings for the DnD-I condition mirrored those from Study 1.

Taking our two experiments together, we provided coherent evidence against a general DnD advantage across three different DnD conditions with a 1-column format. Thus, we believe that the novelty and comparison explanation are ruled out for the time being (see Table 1). But, a different, unknown mechanism has apparently led to an advantage of the DnD-II format. Respondents in this condition made more use of extreme categories. However, this appeared to be an increase in "true" variability rather than ERS variability, which in turn led to slightly higher reliabilities and validities. Nevertheless, the mechanism that led to more variability in category usage remains unclear. It might be the case that the more compact display of categories made extreme categories appear less extreme and therefore more appealing. Or, extreme categories might have been more attractive because less physical effort was required to reach them with the computer mouse.

Even though our results fit into a coherent picture, it is important to point out some limitations of our studies, both in general and in comparison to Böckenholt (2017). First, we had smaller sample sizes, and convenience samples were used. Second, we made all already answered items visible in the DnD-open condition (see Figure 2A) in order to facilitate comparative processes. Post hoc, we realized that this might have been too much information potentially causing a contrary effect. Third, we looked at the effect of response formats from a response style perspective (and investigated reliability and validity, too). However, response formats may have effects in areas not discussed herein, for instance, factor structure, motivation and compliance, or social desirability. These are routes for future research, which may help to scrutinize the DnD format and the accompanying response processes.

Furthermore, it is important to note that implementing a computer-based DnD format involves several challenges: First, a 2-column format makes only sense with an even number of categories. Second, displaying already answered items (on the right-hand side) may take up a lot of space, especially if item texts are long and if the categories are presented in one column. Third, the format is dynamic whereas a Likert-type format is static. Fourth, which and how many items to place on one page may have an even larger impact than in a Likert-type format, where issues like ordering effects have already been documented (e.g., Schwarz, 1999). Fifth, the format is a technical challenge for the respondent (especially for those with visual or motor impairments), and for the researcher/programmer who faces a lot of seemingly arbitrary design choices. Thus, a DnD format can be implemented in many different ways, and it may differ from a Likert-type format in many different ways. In contrast to a DnD format, a Likert-type format—even though it faces several criticism—has the advantage that it is well studied and familiar to many respondents.

Conclusion

Our aim was to investigate whether a drag-and-drop response format may in general reduce the influence of response styles. Our analyses revealed that this is not the case and that the potential advantage may be related to the arrangement of categories. Researchers aiming to make use of a drag-and-drop format should not generally expect positive effects regarding response styles, reliability, or validity, but large negative effects are also not expected. For the time being, we conclude that the Likert-type format is a robust, reliable, well-studied response format that has survived almost a century of constant criticism and probably will live on for decades.

References

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

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

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*, 159–181. doi:10.1111/bmsp.12086

Bolt, D. M. & Johnson, T. R. (2009). Addressing score bias and differential item functioning due to individual differences in response style. *Applied Psychological Measurement*, *33*, 335–352. doi:10.1177/0146621608329891

Bolt, D. M. & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement*, *71*, 814–833. doi:10.1177/0013164410388411

Brown, A. & Maydeu-Olivares, A. (2012). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*, 36–52. doi:10.1037/a0030641

Brummett, B. H., Babyak, M. A., Williams, R. B., Barefoot, J. C., Costa, P. T., & Siegler, I. C. (2006). NEO personality domains and gender predict levels and trends in body mass index over 14 years during midlife. *Journal of Research in Personality*, *40*, 222–236. doi:10.1016/j.jrp.2004.12.002

Couch, A. & Keniston, K. (1960). Yeasayers and naysayers: Agreeing response set as a personality variable. *The Journal of Abnormal and Social Psychology*, *60*, 151–174. doi:10.1037/h0040372

De Beuckelaer, A., Weijters, B., & Rutten, A. (2010). Using ad hoc measures for response styles: A cautionary note. *Quality & Quantity*, *44*, 761–775. doi:10.1007/s11135-009-9225-z

De Boeck, P. & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. doi:10.18637/jss.v048.c01

Duhachek, A. & Iacobucci, D. (2004). Alpha's standard error (ASE): An accurate and precise confidence interval estimate. *Journal of Applied Psychology*, *89*, 792–808. doi:10.1037/0021-9010.89.5.792

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

Henninger, M. & Meiser, T. (2017). *An integration of IRT models for response styles.* Manuscript in preparation.

Hilbig, B. E., Thielmann, I., Hepp, J., Klein, S. A., & Zettler, I. (2015). From personality to altruistic behavior (and back): Evidence from a double-blind dictator game. *Journal of Research in Personality*, *55*, 46–50. doi:10.1016/j.jrp.2014.12.004

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*, 99–114. doi:10.1007/s10869-011-9231-8

Jeon, M. & De Boeck, P. (2016). A generalized item response tree model for psychological assessments. *Behavior Research Methods*, *48*, 1070–1085. doi:10.3758/s13428-015-0631-y

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*, 161–177. doi:10.1080/00273171.2013.866536

Lee, K. & Ashton, M. C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment*, *18*, 182–191. doi:10.1037/1040-3590.18.2.182

Machunsky, M. & Meiser, T. (2006). Personal need for structure als differenzialpsychologisches konstrukt in der sozialpsychologie [Personal need for structure as a construct of dispositional differences in social psychology]. *Zeitschrift für Sozialpsychologie*, *37*, 87–97. doi:10.1024/0044-3514.37.2.87

McKeown, B. & Thomas, D. B. (1988). *Q methodology.* Thousand Oaks, CA: Sage.

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455. doi:10.1037/a0028085

Moshagen, M., Hilbig, B. E., & Zettler, I. (2014). Faktorenstruktur, psychometrische Eigenschaften und Messinvarianz der deutschsprachigen Version des 60-Item HEXACO Persönlichkeitsinventars [Factor structure, psychometric properties, and measurement invariance of the German-language version of the 60-item HEXACO personality inventory]. *Diagnostica*, *60*, 86–97. doi:10.1026/0012-1924/a000112

Muthén, L. K. & Muthén, B. O. (2012). *Mplus. Statistical analysis with latent variables, Version 7.* Los Angeles, CA: Muthén & Muthén.

Neuberg, S. L. & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology*, *65*, 113–131. doi:10.1037/0022-3514.65.1.113

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*, 867–872. doi:10.1016/j.jesp.2009.03.009

Plieninger, H. (2016). Mountain or molehill? A simulation study on the impact of response styles. *Educational and Psychological Measurement*, *77*, 32–53. doi:10.1177/0013164416636655

Plieninger, H. & Heck, D. W. (2017). *A new model for acquiescence at the interface of psychometrics and cognitive psychology*. Manuscript submitted for publication.

Plieninger, H. & Meiser, T. (2014). Validity of multiprocess IRT models for separating content and response styles. *Educational and Psychological Measurement*, *74*, 875–899. doi:10.1177/0013164413514998

R Core Team. (2017). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org

Revelle, W. (2017). psych: Procedures for psychological, psychometric, and personality research (Version 1.7.5). Retrieved from https://CRAN.R-project.org/package=psych

Robitzsch, A., Kiefer, T., & Wu, M. (2017). TAM: Test analysis modules (Version 2.8-12). Retrieved from https://github.com/alexanderrobitzsch/TAM

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Society. Richmond, VA. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Savalei, V. & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*, 407–424. doi:10.1080/00273171.2014.931800

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*, 93–105. doi:10.1037/0003-066X.54.2.93

Sutin, A. R., Ferrucci, L., Zonderman, A. B., & Terracciano, A. (2011). Personality and obesity across the adult life span. *Journal of Personality and Social Psychology*, *101*, 579–592. doi:10.1037/a0024286

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *33*, 529–554. doi:10.1086/214483

Thurstone, L. L. (1930). A scale for measuring attitude toward the movies. *The Journal of Educational Research*, *22*, 89–94. doi:10.1080/00220671.1930.10880071

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*, 195–217. doi:10.1093/ijpor/eds021

von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2017). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement*. Advance online publication. doi:10.1177/0146621617730389

Weijters, B., Cabooter, E. F. K., & 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*, 236–247. doi:10.1016/j.ijresmar.2010.02.004

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods*, *15*, 96–110. doi:10.1037/a0018721

Weng, L.-J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, *64*, 956–972. doi:10.1177/0013164404268674

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC International Handbook of Testing and Assessment* (pp. 349–363). doi:10.1093/med:psych/9780199356942.003.0024

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*, 304–324. doi:10.1177/0013164415591848

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

Wetzel, E., Roberts, B., Fraley, R., & Brown, A. (2016). Equivalence of narcissistic personality inventory constructs and correlates across scoring approaches and response formats. *Journal of Research in Personality*, *61*, 87–98. doi:10.1016/j.jrp.2015.12.002

Wickham, H. (2017). tidyverse: Easily install and load “tidyverse” packages (Version 1.1.1). Retrieved from https://CRAN.R-project.org/package=tidyverse

Zhao, K. & Smillie, L. D. (2015). The role of interpersonal traits in social decision making. *Personality and Social Psychology Review*, *19*, 277–302. doi:10.1177/1088868314553709

Appendix A

Descriptive Statistics of Core Variables of Study 1 and 2

Descriptive statistics, namely, M, SD, and correlations, of core variables of Study 1 are summarized in Table A1, and variables of Study 2 are summarized in Table A2. The tables contain the values for the whole samples as well as values for each condition separately. Note that five response categories were used in Study 1, whereas six categories were used in Study 2. Thus M and SD for honesty-humility are not directly comparable across studies. Moreover, the heterogeneous items that were used to measure MRS, ERS, and ARS are also not directly comparable, because they are based on different coding schemes. For example, MRS indicators for 5-point items were build using a coding scheme of $(0, 0, 1, 0, 0)$, whereas a scheme of $(0, 0, 1, 1, 0, 0)$ was used for 6-point items.

Table A1: Study 1 (5-point scale): Means, Standard Deviations, and Correlations

	Variable	M	SD	Correlations					
Study 1 ($n = 644$)				1	2	3	4	5	6
1	Honesty-humility	3.59	0.60						
2	Conscientiousness	3.68	0.60	.06					
3	MRS	0.27	0.10	.08	.03				
4	ERS	0.48	0.11	−.05	−.04	−.83			
5	ARS	0.39	0.11	−.28	.16	−.58	.50		
6	Dictator Game	4.13	2.00	.28	−.04	.07	−.08	−.12	
7	Body mass index	23.11	4.21	.07	−.04	−.02	.05	−.02	−.02
Likert ($n = 203$)									
1	Honesty-humility	3.54	0.57						
2	Conscientiousness	3.66	0.63	.02					
3	MRS	0.27	0.10	.15	.01				
4	ERS	0.47	0.11	−.10	−.03	−.82			
5	ARS	0.38	0.11	−.26	.17	−.63	.52		
6	Dictator Game	4.15	2.03	.26	−.05	.04	−.08	−.01	
7	Body mass index	23.11	4.13	.14	−.13	−.04	.10	.00	.01
DnD-open ($n = 218$)									
1	Honesty-humility	3.58	0.61						
2	Conscientiousness	3.65	0.59	.06					
3	MRS	0.27	0.10	−.02	.08				
4	ERS	0.48	0.11	.02	−.08	−.86			
5	ARS	0.40	0.11	−.29	.13	−.60	.52		
6	Dictator Game	4.02	2.00	.27	.06	.08	−.12	−.13	
7	Body mass index	23.14	4.38	.08	−.01	−.07	.10	−.06	−.02
DnD-shut ($n = 223$)									
1	Honesty-humility	3.65	0.62						
2	Conscientiousness	3.74	0.59	.08					
3	MRS	0.26	0.09	.13	.02				
4	ERS	0.50	0.10	−.10	−.02	−.83			
5	ARS	0.39	0.11	−.31	.18	−.53	.45		
6	Dictator Game	4.22	1.99	.32	−.12	.10	−.06	−.20	
7	Body mass index	23.08	4.14	.00	.01	.05	−.04	.00	−.04

Note. MRS, ERS, and ARS are response style measures based on 39 heterogeneous items.

Table A2: Study 2 (6-point scale): Means, Standard Deviations, and Correlations

	Variable	M	SD	Correlations				
Study 2 ($n = 506$)				1	2	3	4	5
1	Honesty-humility	4.41	0.70					
2	Personal Need for Structure	3.84	0.69	−.05				
3	MRS	0.47	0.17	−.03	−.01			
4	ERS	0.35	0.14	.03	.04	−.94		
5	ARS	0.54	0.11	−.23	.10	.01	−.02	
6	Dictator Game	4.48	1.85	.23	−.02	.08	−.10	−.03
Likert ($n = 181$)								
1	Honesty-humility	4.44	0.66					
2	Personal Need for Structure	3.78	0.68	.00				
3	MRS	0.48	0.18	−.01	.08			
4	ERS	0.34	0.15	.00	−.05	−.94		
5	ARS	0.53	0.11	−.21	.06	.07	−.09	
6	Dictator Game	4.59	1.71	.23	.08	.09	−.08	−.09
DnD-I ($n = 162$)								
1	Honesty-humility	4.42	0.67					
2	Personal Need for Structure	3.87	0.63	−.03				
3	MRS	0.49	0.16	−.03	−.01			
4	ERS	0.34	0.13	.07	.04	−.96		
5	ARS	0.54	0.11	−.24	.18	−.02	−.02	
6	Dictator Game	4.62	1.77	.14	.02	.09	−.09	.09
DnD-II ($n = 163$)								
1	Honesty-humility	4.36	0.77					
2	Personal Need for Structure	3.87	0.75	−.11				
3	MRS	0.44	0.16	−.06	−.09			
4	ERS	0.38	0.14	.04	.14	−.94		
5	ARS	0.54	0.10	−.26	.09	−.05	.06	
6	Dictator Game	4.21	2.06	.29	−.12	.03	−.10	−.07

Note. MRS, ERS, and ARS are response style measures based on 31 heterogeneous items.

Appendix B

Response Frequencies in Data From Böckenholt (2017)

As described above, Böckenholt (2017) administered the 12 PNS items with six response categories and had roughly 500 participants in each of three conditions. The Likert condition used a standard rating format and the DnD condition is displayed in Figure 1. In a third condition (Funnel) not discussed herein, participants first indicated agreement versus disagreement on a two-point scale and then rated the intensity of this belief on a three-point scale. Data for 11 PNS items are available from the journal's website and the analyses revealed the following results: Likert condition had $\alpha = .66$, $M = 3.53$, and $SD = 0.56$; DnD condition had $\alpha = .79$, $M = 3.55$, and $SD = 0.72$; Funnel condition had $\alpha = .68$, $M = 3.55$, and $SD = 0.62$. The response distributions in the three conditions shown in Figure B1 revealed a specific response pattern in the Likert condition, where the two intermediate categories were selected with a frequency of 71 %.

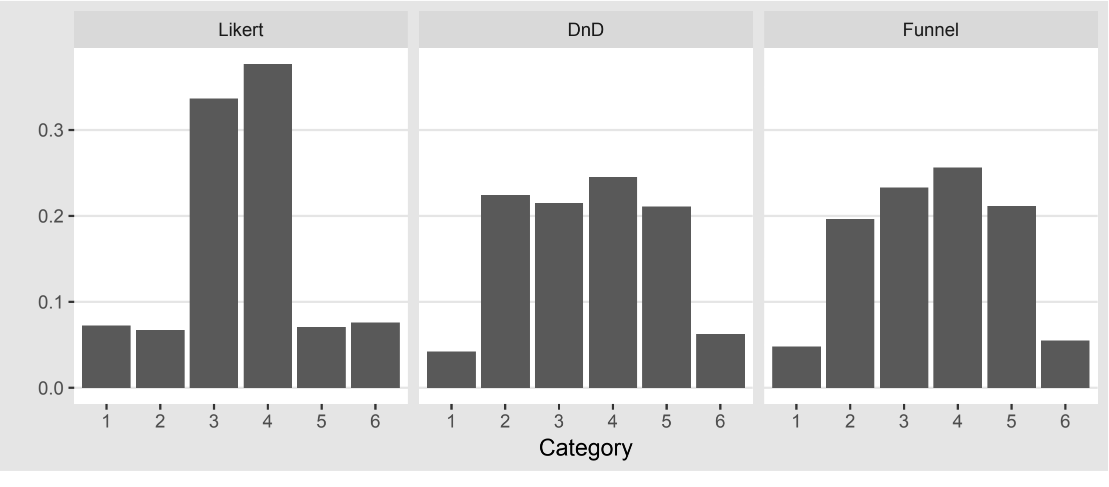


FIGURE B1: Barplots of relative frequencies observed across all items and participants in each of the three conditions in Böckenholt (2017).