Doctoral Dissertation

Disentangling Substantive Traits and Faking in High-Stakes Personality Assessments Using Item Response Theory Modeling

Timo Seitz

Inaugural dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Social Sciences in the DFG Research Training Group “Statistical Modeling in Psychology” (SMiP) at the University of Mannheim

Dean of the School of Social Sciences:
Dr. Julian Dierkes

1st Supervisor:
Prof. Dr. Thorsten Meiser

2nd & 3rd Supervisors:
Prof. Dr. Eunike Wetzel (RPTU University Kaiserslautern-Landau)
Prof. Benjamin E. Hilbig, Ph.D. (RPTU University Kaiserslautern-Landau)

Dissertation Evaluators:
Prof. Dr. Florian Keusch
Prof. Dr. Steffi Pohl (Freie Universität Berlin)

Disputation Date:
March 3, 2026

Table of Contents

Attachments:

Acknowledgements

Without the extensive support of several people, this dissertation would not have been possible in its final form. I would especially like to express my gratitude to …

- … Thorsten Meiser for being a great primary advisor. His excellent lectures once made me decide for a PhD in psychometrics, and his way of approaching research questions and intellectual problems has inspired me ever since. As a supervisor, he gave me the freedom to develop and pursue own ideas while challenging my work whenever necessary, thus providing all resources for productive and good research.
- … Eunike Wetzel and Benni Hilbig for their supportive roles as secondary SMiP advisors, and particularly for their valuable contributions regarding Article I.
- … Emre Alagöz for his support with Article II. Specifically, I am thankful for sharing his expertise and model syntax on mixture modeling, without which the projects would have taken several months longer.
- … Esther Ulitzsch for inviting me to the University of Oslo to work together on the model for Article III. This was certainly one of the highlights of my PhD.
- … Florian Keusch and Steffi Pohl for their willingness to act as evaluators of this dissertation.
- … Maik Spengler for giving me the chance to use real high-stakes data for my research right from the start of our collaboration. Especially, he was the leading force behind the collection of item-level response times, without which Article III would not have been possible. Likewise, I am grateful for the many valuable insights from a practitioner's point of view on diagnostics, helping me not to lose sight of this important aspect. I would also like to thank all other colleagues from the Research & Development team at HR Diagnostics, who made my freelance work in the testing industry very enjoyable.
- … Dietrich Wagener for inspiring me about statistics and test theory in my first bachelor semesters. Without this early interest in methods and psychological assessment, I would have never come to do a PhD in this field.
- … Martin Brandt for enthusiastically teaching *R* and thus convincing me of its great potential.
- … my chair mates Vincent, Emre, Selina, Hannah, Julia, Martin, Viola, Nils, Fabiola, and Marcel as well as all other colleagues from fifth floor in L13 for all the fun lunch meetings, after-work beers, and private activities. This made work not feel like work.
- … all fellow SMiPsters for creating a supportive and motivating environment during the qualification program. I also want to thank Anke and Annette for organizing these nice SMiP events, as well as Silke for the affirmative coaching sessions.
- … Berit for just being there.
- … my parents Sonja and Thomas for supporting me with everything from day one.

Summary

Self-report questionnaires are frequently used to measure psychological constructs like personality traits, interests, and attitudes. When such questionnaires are employed in high-stakes contexts, where assessment results have important consequences for test-takers, there is a risk that test-takers deliberately distort their responses according to what is socially desirable. This response bias is known as faking, and has been studied by social scientists and research methodologists for decades. Most of the approaches that have been developed either aim at the detection or prevention of faking. In this dissertation, however, I address the response bias of faking by means of latent variable models. Such models allow disentangling different influences on item responses, thus enabling both a statistical control of faking and a substantive investigation of the response process associated with faking.

In a series of three articles, I develop and examine psychometric models to separate the measurement of substantive traits from faking, and meet the issue from an item and test construction perspective. The models of all three articles are based on the multidimensional nominal response model (MNRM), which is a flexible item response theory (IRT) model that allows to address limitations of previous latent variable models of faking. Article I (Seitz et al., 2024) presents an examination of conditions under which the MNRM can effectively adjust substantive trait scores and latent correlations for faking, particularly concerning the question of how item content needs to be related to social desirability for the model to perform well. Article II (Seitz, Alagöz, & Meiser, 2025) extends the MNRM in a mixture modeling framework, assuming that test-takers differ qualitatively in the response strategy they employ. Article III (Seitz & Ulitzsch, 2025) introduces a further model extension, which makes use of item-level response times (RT) and allows for switches between response strategies over the course of the questionnaire. Along with simulations to study statistical properties of the models, all articles also include empirical demonstrations in experimental questionnaire data or real high-stakes datasets from personnel selection.

The conducted research shows that faking is not an intractable response bias but that the application of sophisticated psychometric models can help to better manage its adverse effects. Beyond methodological contributions, the models of this dissertation also provide a framework researchers can use to enhance the understanding of the substantive nature of faking. Nevertheless, future studies should provide more validation evidence, especially with regard to the added value of the models in applied diagnostic contexts.

Articles

This cumulative dissertation thesis is the result of research projects I have conducted as part of my associate membership in the DFG Research Training Group "Statistical Modeling in Psychology" (SMiP).[*] It is based on the following three articles, two of which have already been published in peer-reviewed journals and one has been submitted for publication:[†]

ARTICLE I:

Seitz, T., Wetzel, E., Hilbig, B. E., & Meiser, T. (2024). Using the multidimensional nominal response model to model faking in questionnaire data: The importance of item desirability characteristics. *Behavior Research Methods, 56*(8), 8869–8896. https://doi.org/10.3758/s13428-024-02509-x

ARTICLE II:

Seitz, T., Alagöz, Ö. E. C., & Meiser, T. (2025). Disentangling qualitatively different faking strategies in high-stakes personality assessments: A mixture extension of the multidimensional nominal response model. *Educational and Psychological Measurement, 85*(6), 1237–1277. https://doi.org/10.1177/00131644251341843

ARTICLE III:

Seitz, T., & Ulitzsch, E. (2025). *Faking in high-stakes personality assessments: A response-time-based latent response mixture modeling approach* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

In the following synopsis, I will first introduce the topic of faking in high-stakes personality assessments and elaborate on previous psychometric modeling approaches. I will then summarize the three articles forming the core of my dissertation (the full texts are attached to the synopsis). Last, I will integrate the findings, discuss implications for research and practice, and derive directions for future work.

[*] This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2277 – Projektnummer 310365261.

[†] After handing in the dissertation but before defending it, a revised version of the manuscript of Article III has been accepted for publication in *Educational and Psychological Measurement*.

1 Introduction

In psychology and other social sciences, most constructs of interest are not directly observable. To measure these constructs, researchers and practitioners have to make use of psychological tests. Many psychological constructs, including personality traits, interests, and attitudes, are thereby assessed through self-report questionnaires. Such questionnaires typically consist of a set of statements (i.e., items) to which test-takers respond using a rating scale with a certain number of graded agreement categories. In fact, a plethora of research has shown that constructs measured through self-report questionnaires employing rating scales consistently predict variables like prosocial behavior (e.g., Habashi et al., 2016), risky health actions (e.g., Atherton et al., 2014), or the susceptibility to psychological disorders (e.g., Kotov et al., 2010). Also, in the context of industrial and organizational psychology, self-reported personality traits have been found to exhibit meaningful associations with job satisfaction (e.g., Judge et al., 2002) and job performance (e.g., Barrick & Mount, 1991; Salgado, 2003; Shaffer & Postlethwaite, 2012). In particular, relationships with performance outcomes have led to the use of self-report personality tests for hiring decisions and other selection purposes (Diekmann & König, 2015; Nikolaou & Foti, 2018; Ones et al., 2007).[1]

Despite the reported predictive validity of self-report questionnaires, these measures are not free from systematic biases. In the psychometric literature, it is well known that responses to rating scale items can be biased by different response tendencies (Paulhus, 1991). Such response biases can be divided into response styles and response sets (Jackson & Messick, 1958). Response styles are tendencies of test-takers to prefer certain kinds of rating scale categories irrespective of the item content (Baumgartner & Steenkamp, 2001; Van Vaerenbergh & Thomas, 2013). These include extreme response style (ERS; tendency toward extreme response categories), midscale response style (MRS; tendency toward midpoint categories), and acquiescent response style (ARS; tendency to generally agree with statements). Response styles have been shown to be consistent over time (e.g., Weijters et al., 2010; Wetzel et al., 2016) and across the measurement of different traits (e.g., Austin et al., 2006; Wetzel et al., 2013), thus representing stable interindividual difference variables. In contrast, response sets reflect tendencies of test-takers that are inherent to the item content and the specific assessment context. Examples are frame-of-reference effects (e.g., Schmit et al., 1995) and specific forms of careless responding (e.g., Meade & Craig, 2012). The most prominent response set, however, is *socially desirable responding* (SDR; Paulhus, 2002), which is the tendency to describe

[1] Throughout this dissertation, I use the terms "test" and "questionnaire" interchangeably when referring to self-report instruments for assessing personality.

oneself in an overly positive light. Particularly in contexts where assessment results have important consequences for test-takers, this response bias can be expected to have pronounced effects. A prominent example of such high-stakes contexts are selection settings, such as job applications or college admissions.

1.1 Faking as a Form of Socially Desirable Responding (SDR)

According to Paulhus' (1984, 2002) framework, SDR can have two different forms, namely a self-directed (*self-deception*) and an other-directed form (*impression management*). Self-deception is the tendency to enhance the private self-image, which can have agentic and communal components. In contrast, impression management is the tendency to deliberately create a favorable self-presentation toward others, which can also have agentic and communal components. It usually occurs in high-stakes assessment contexts and is commonly referred to as *faking*. This dissertation focuses on the deliberate form of SDR, that is, faking. As described in Section 1.1.1, faking can have severe adverse effects on psychological assessment, especially in high-stakes contexts, creating a strong need for methods that can account for it.

Ziegler et al. (2011, p. 8) formulated the following working definition of faking: "Faking represents a response set aimed at providing a portrayal of the self that helps a person to achieve personal goals. Faking occurs when this response set is activated by situational demands and person characteristics to produce systematic differences in test scores that are not due to the attribute of interest." Several theoretical models on the process of faking have spelled out antecedents of faking behavior (e.g., Goffin & Boyd, 2009; Roulin et al., 2016; Snell et al., 1999; Ziegler, 2011). What is common to the different theoretical models is that faking constitutes a complex interplay of person and situation effects, including ability, opportunity, as well as motivation and intention to fake (Dunlop et al., 2022; Ellingson & McFarland, 2011; McFarland & Ryan, 2006; Mueller-Hanson et al., 2006; Tett et al., 2012; Tett & Simonet, 2011). Hence, faking cannot be considered a stable person characteristic like response styles, but rather a situation-specific person variable that systematically biases responses to self-report items.

1.1.1 Effects of Faking

Faking has numerous adverse effects on the measurement of psychological constructs (Ziegler et al., 2011). The most robust effect is that it biases mean scores of items and scales. In particular, it increases mean scores on desirable traits and decreases mean scores on undesirable traits. This has been shown in experimental studies (Viswesvaran & Ones, 1999) as well as in studies comparing data from job application and low-stakes contexts (Birkeland et al., 2006;

Hu & Connelly, 2021). Effect sizes vary between studies and operationalizations, but findings from a recent meta-analysis on this topic (based on within-subjects studies comparing data from actual high-stakes and low-stakes contexts) indicate that faking changes mean scores by at least half a standard deviation (Hu & Connelly, 2021). Increased or decreased mean scores were per se not problematic if all test-takers shifted their responses between low-stakes and high-stakes settings to the same extent. In this case, rank orders of test-takers would be retained, and one would only need suitable standardization samples to interpret test scores appropriately. However, mean shifts usually go along with ceiling or floor effects and hence a reduced variability of scores (Hu & Connelly, 2021). In addition, test-takers in fact do differ in the extent to which they shift their responses between low-stakes and high-stakes settings. This is evidenced by correlations between low-stakes and high-stakes conditions that are typically lower than test-retest correlations in regular research settings (Hu & Connelly, 2021; Krammer et al., 2017). Also, it has been found that some but not all job applicants really engage in faking. For instance, based on the randomized response technique, Donovan et al. (2003) estimated that 56% of job applicants exaggerate their positive attributes and 17% make outright false claims about themselves, which is somewhat in line with the estimated general faking prevalence of 30% (± 10%) in Griffith and Converse's (2011) literature review. Regardless of the precise number, all findings in this context indicate heterogeneity in the response behavior in high-stakes personality assessments, leading to biased rank orders of test-takers due to faking.

Biased rank orders can have multiple consequences: First, they are an issue of construct validity. If there is an additional source of systematic variance, the measurement of the construct becomes less pure. From a psychometric perspective, contaminated measurement is in itself undesired and calls for countermeasures. Second, in applied settings like personnel selection contexts, biased rank orders imply that different persons will be selected based on the test scores. Especially when selection ratios are small, those who are selected will mostly be those who have faked their responses (Mueller-Hanson et al., 2003). Third, biased rank orders can lead to different correlations with criterion variables (e.g., Komar et al., 2008; Paunonen & LeBel, 2012). Recent meta-analyses have found generally lower criterion-related validities of personality measures in high-stakes than in low-stakes contexts (Loy et al., 2025; Speer et al., 2025). However, there are also studies finding rather limited effects of faking on how well outcomes can be predicted based on personality test scores (e.g., Ones et al., 2007; Paunonen & LeBel, 2012).

Additionally, faking has been shown to bias inter-item and inter-scale correlations. Specifically, by introducing another source of systematic variance, faking creates an additional dimension in a test's factor structure, which Schmit and Ryan (1993) termed the "ideal-employee factor" (see also Klehe et al., 2012). When not accounting for it, correlations between items as well as correlations between scales are inflated (e.g., Ellingson et al., 1999; Christiansen et al., 2021). For instance, in a sample of job applicants, Seitz, Spengler, and Meiser (2025) found intercorrelations between the Big Five personality factors ranging from .52 to .81, whereas a meta-analysis on Big Five intercorrelations in research settings yielded values ranging from .17 to .43 (van der Linden et al., 2010). Hence, faking undermines construct validity also in terms of diminished discriminant validity, and thus impedes nuanced personality profiles across different traits.

1.1.2 Different Perspectives on Faking

The above-described psychometric effects of faking are mostly agreed upon (Ziegler, 2011). Based on these effects, the conclusion seems warranted that faking is something bad and requires measures to counter it ("faking-is-bad" (FIB) perspective). However, some researchers and practitioners also argue that faking is something good and represents itself a desirable attribute ("faking-is-good" (FIG) perspective; see Tett & Simonet, 2021, for a juxtaposition of the two perspectives). Proponents of the FIG perspective (e.g., Hogan & Blickle, 2013; Marcus, 2009) see faking as a sign of social skills and a demonstration of effective self-presentation. They argue that people successful in managing their impression when responding to a questionnaire in a job application will also be successful in managing their impression and behavior on the job, that is, will successfully navigate the social world. Such a perspective on faking is related to research on people's ability to identify criteria (ATIC), which is a construct defined as the ability to correctly perceive performance criteria in evaluative contexts and has been found to be associated with intelligence and actual performance (Klehe et al., 2012; Kleinmann et al., 2011). Also, FIG proponents refer to the occasional finding that controlling for SDR and faking in personality measures reduces predictive validity (Li & Bagger, 2006) because, as they argue, essential (and predictive) components of personality are partialed out. At least, they are convinced that faking can generally be ignored (Morgeson et al., 2007) or does not need to be countered (Marcus, 2022).

Whereas the FIG perspective is grounded on socioanalytic (Hogan & Blickle, 2013), self-presentational (Marcus, 2009), and pragmatic considerations, the FIB perspective adopts a strong psychometric point of view (e.g., Furr, 2021; Tett & Simonet, 2021), which sees construct validity as the most important property of measurement. According to the current unitarian understanding (Binning & Barrett, 1989; Sireci, 2009), validity represents "the degree to which evidence and theory support the interpretation of test scores proposed by the test user" (American Educational Research Association et al., 2014, p. 11). Crucially, however, the "proposed interpretation includes specifying the construct the test is intended to measure". Validity of a test is hence the degree to which the test measures what it is supposed to measure. Important is the specification of a targeted construct, which rules out a merely empiricist view on validity where a significant association between a test and a criterion renders a test "valid". This conception of validity implies that faking is detrimental to validity because it introduces systematic bias to item responses such that test scores do not only reflect the construct of interest (i.e., the substantive trait). Hence, proponents of the FIB perspective see faking as something that needs to be addressed.

In the present dissertation, I uphold the FIB perspective and agree with Tett and Simonet's (2021, p. 6) claim that "the FIG perspective is psychometrically flawed and counterproductive to personality-based selection targeting trait-based fit". Nevertheless, I want to emphasize that adopting the FIB perspective does not imply that accounting for faking is in any way beneficial for construct validity and the prediction of outcome variables like job performance. I will come back to this issue in Section 6.3.

1.1.3 Approaches to Dealing With SDR and Faking

Following the FIB perspective, there have been two main lines of research dealing with SDR and faking, which either focus on its detection or its prevention (see Dunlop et al., 2025, for a recent review; Fan et al., 2012). None of the approaches, however, comes without drawbacks or is applicable in every context.

Detection. One line of research aims at detecting (or, more generally, measuring) SDR and faking. An example of a direct faking detection method are SDR scales or other so-called validity scales (see Paulhus, 2002; Paulhus & Trapnell, 2008, for overviews), which should quantify each test-taker's degree of SDR and faking. These scales typically contain items about desirable attributes that are effectively very rare (e.g., "I have never lied to anybody") as well as items about undesirable attributes that are in fact very common (e.g., "I occasionally make silly mistakes"). Test-takers endorsing many of the former and rejecting many of the latter items receive high scores on such scales. For decades, SDR scales have been frequently used. However, a striking limitation of these scales is that they confound SDR and faking with substantive trait variance (e.g., de Vries et al., 2014; McCrae & Costa, 1983). This is reflected in moderate to strong correlations of SDR scales with the Big Five (Li & Bagger, 2006; Ones et al., 1996). To adjust substantive trait scores for SDR and faking, it is hence inappropriate to partial SDR scale scores from substantive trait scale scores, as this removes a considerable proportion of substantive variance from trait scores (Griffith & Peterson, 2008; Reeder & Ryan, 2011).

As opposed to direct faking detection approaches, there are also indirect detection techniques. These methods do not require a designated instrument administered along with the actual personality questionnaire, but make use of the given responses themselves or collateral data (see Goldammer et al., 2024, for an overview). Most of these methods are based on response patterns. For example, there are measures of response homogeneity (e.g., the item-level covariance index; Christiansen et al., 2017), measures of response inconsistency (e.g., person-fit indices in item response theory (IRT) models; LaHuis & Copeland, 2009; Zickar & Drasgow, 1996), and measures scrutinizing responses to specific items (e.g., the idiosyncratic item response method; Kuncel & Borneman, 2007). Other methods rely on incidental trace data, such as mouse clicking behavior, eye movements, or response latencies (e.g., Kuric et al., 2025; Mazza et al., 2024).

Prevention. In contrast to aiming at the detection of faking, the other line of research seeks to prevent faking at data collection. That is, the goal is to have data that is free of faking in the first place, rather than having to deal with it post hoc. One approach that has become increasingly popular over the last years is the multidimensional forced-choice (MFC) response format (see Lee et al., 2025, for an overview). In MFC, test-takers do not respond to items using a classical rating scale, but are confronted with blocks of two or more items from multiple dimensions and have to rank the items according to how well the items describe their personality. Conceptually, faking and SDR should be prevented if items within blocks are matched according to social desirability, since test-takers cannot endorse all desirable and reject all undesirable items but are forced to make a choice among items of equal desirability. Several meta-analyses have investigated the susceptibility of the MFC format to faking when items within blocks are desirability-matched (e.g., Cao & Drasgow, 2019; Martínez & Salgado, 2021; Speer et al., 2023). In short, MFC measures seem to be less fakable than rating scale measures in terms of mean differences between low-stakes and high-stakes conditions, exhibit strong convergent correlations with rating scale measures, and have overall similar criterion-related validity (see also Wetzel & Frick, 2020; Wetzel et al., 2021). However, they are not fully faking-proof and are associated with several challenges: First, MFC tests come with considerable test construction efforts, especially in terms of piloting items with regard to desirability and matching them to create appropriate blocks (see Brown et al., 2017; Li et al., 2025). Second, the response format featuring comparative judgments increases the cognitive effort for test-takers (Brown & Bartram, 2009; Sass et al., 2020) and introduces other potential biases as items are placed in a strong contextual framework within blocks (see Frick, 2022; Fuechtenhans & Brown, 2023). Third, MFC tests often have a reliability that is too low for individual diagnostic purposes (Bürkner et al., 2019; Schulte et al., 2021). Fourth, when scored with classical methods, MFC tests yield ipsative test scores (Brown, 2010), that is, scores that can only be interpreted within persons but not between persons. With the application of Thurstonian item response theory (TIRT) modeling, however, it is possible to obtain normative scores from MFC measures (Brown & Maydeu-Olivares, 2011, 2013). Full normativity is nevertheless hard to achieve even with TIRT (see Schünemann, 2025). In particular, having both positively- and negatively-keyed items within blocks seems to be crucial for obtaining normative scores (Bürkner et al., 2019; Frick et al., 2023). This design feature of mixed-keyed item blocks, however, provides a big challenge for composing item blocks with matched desirability (cf. Section 6.3).

Along with MFC, there are also approaches that seek to prevent faking in rating scale measures by refining item content. Bäckström et al. (2009; see also Bäckström & Björklund, 2024; Bäckström et al., 2023), for instance, proposed to neutralize the valence of items to make them less susceptible to SDR and faking. The authors indeed found that such item refinements can reduce the variance attributable to a general SDR/faking factor while retaining the loading structure of the substantive traits and preventing mean score inflation. However, this approach as well as a related method (Widhiarso et al., 2019) still await widespread application (see Wood et al., 2022, 2024, for an exception), and are also associated with considerable test construction efforts. Particularly, it can be challenging or even impossible to create items with neutral valence for inherently desirable or undesirable traits (Wood et al., 2022). In such cases, neutralizing item content can change the meaning of the assessed construct.

Furthermore, other approaches aiming at the prevention of faking have been developed in recent years, including gamified assessments (e.g., Landers & Sanchez, 2022; Nikolaou & Katsadoraki, 2025), warnings about implemented faking detection measures (e.g., Feeney et al., 2023; Moon et al., 2025), implicit personality tests (e.g., Cook et al., 2024), as well as rapid response measurement (Meade et al., 2020). These methods, however, all involve psychometric and practical tradeoffs, such as limited convergent validities or questionable generalizability of effects (see Dunlop et al., 2025).

1.2 Psychometric Modeling Approaches to Faking

All approaches outlined in the previous section either aim at detecting or preventing faking. Each approach has its own merits, however, there are general limitations associated with detection and prevention methods. With regard to faking detection methods, these approaches only yield a measurement of faking, either in terms of a continuous quantification or a binary classification. However, even if the measurement and eventually also the detection were successful, faking would not be readily controlled for in the sense that test-takers' trait scores are adjusted for faking. Researchers and practitioners would only be left with a piece of information about the trustworthiness of the data from individual test-takers, and would have to decide for themselves how to proceed with the respective data. Regarding faking prevention methods, these approaches imply non-negligible test construction efforts and, by definition, are not applicable when data has already been collected. Also, because faking is eliminated from the outset, faking prevention methods do not allow for an investigation of the response process associated with faking. Understanding the nature of faking is, however, essential to be able to counter it in a sophisticated manner, and it is of interest from a mere research point of view (cf. Bensch et al., 2019).

A different approach is to use psychometric latent variable models. Such models aim at disentangling the simultaneous influence of substantive traits and response biases (i.e., faking) on item responses, and are thus not limited to either the detection or the prevention of faking. Instead, they yield faking-adjusted substantive trait scores as well as a measurement of each test-taker's degree of faking, making them useful for both diagnostic and substantive research purposes. Most of the existing latent variable models of faking are applicable to Likert rating scale data, which is still the most frequently used response format in self-report personality questionnaires (Jebb et al., 2021).

1.2.1 Previous Latent Variable Models of Faking

A straightforward latent variable model to account for faking is the bifactor model (e.g., Hendy et al., 2021; see also Klehe et al., 2012; Schmit & Ryan, 1993). The bifactor model of faking is a structural equation model (SEM) in which items are specified to load on their respective substantive trait factor (i.e., a specific factor) as well as on a global faking factor (i.e., the general factor), while all factors are restricted to be orthogonal. Previous work fitting such a model to personality data has found that substantive trait scores based on specific factors are less distorted by faking than classical scale scores (Hendy et al., 2021). Also, the general faking factor has been found to be related to external covariates (Klehe et al., 2012). However, there are several limitations associated with the bifactor model approach: First, the model can easily be empirically underidentified (Podsakoff et al., 2003). For instance, when only one substantive trait is modeled, the model is underidentified and reduces to an exploratory two-factor model. Second, the assumption of orthogonal factors is oftentimes unreasonable, especially with regard to specific factors that should capture truly correlated personality traits. Third, because of the restriction of orthogonal factors, specific factors in bifactor models reflect residual factors that represent what is left over after accounting for the variance of the general factor. Hence, given common variance among substantive traits (as is the case for the Big Five; van der Linden et al., 2010), a general faking factor captures substantive trait variance such that specific factors can no longer be interpreted as manifestations of the actual personality traits (cf. Chen et al., 2012; Koch et al., 2018).

Another modeling technique used to account for faking is finite mixture modeling (e.g., Leite & Cooper, 2010; Zickar et al., 2004), in which the data is assumed to consist of distinct latent classes representing different types of response behavior. Zickar et al. (2004), for instance, found three latent classes when applying a mixture IRT model to personality data from a sample of job applicants. The classes were characterized by different ordering and spacing of threshold parameters, leading the authors to interpret the classes as an honest, slight-faking, and extreme-faking class. However, such a mixture modeling approach is purely exploratory and atheoretical, leaving room for interpretation concerning the meaning of classes. Also, the observed classes do not necessarily capture response tendencies related to faking, as they may well represent other response biases, such as response styles (e.g., Böckenholt & Meiser, 2017), or obscure mixings of faking with other sources of heterogeneity. Ziegler et al. (2015) combined the approach of mixture modeling with latent change score modeling to acknowledge that faking does not have to be a discrete variable and to facilitate the interpretability of results. This method, however, comes with the limitation that it requires data from the same test-takers under both a high-stakes and low-stakes condition, which is not realizable in most applied measurement contexts.

A more promising psychometric approach is to use multidimensional IRT modeling. In such models, faking is represented as an additional latent dimension along with substantive trait dimensions. Some studies have used item response tree (IRTree) models in this context (LaHuis et al., 2019; Lee et al., 2022; Sun et al., 2022). These models regard a response as a product of multiple decision-making processes, where each process can be influenced by different latent dimensions. For example, Sun et al. (2022) used a model in which the decision of generally agreeing with an item is associated with the substantive trait whereas the decision of choosing extreme response categories is associated with a latent faking variable. Even though the faking variable significantly predicted whether test-takers were job applicants or job incumbents in the study by Sun et al. (2022), IRTree models of faking cannot readily separate faking from response styles like ERS.

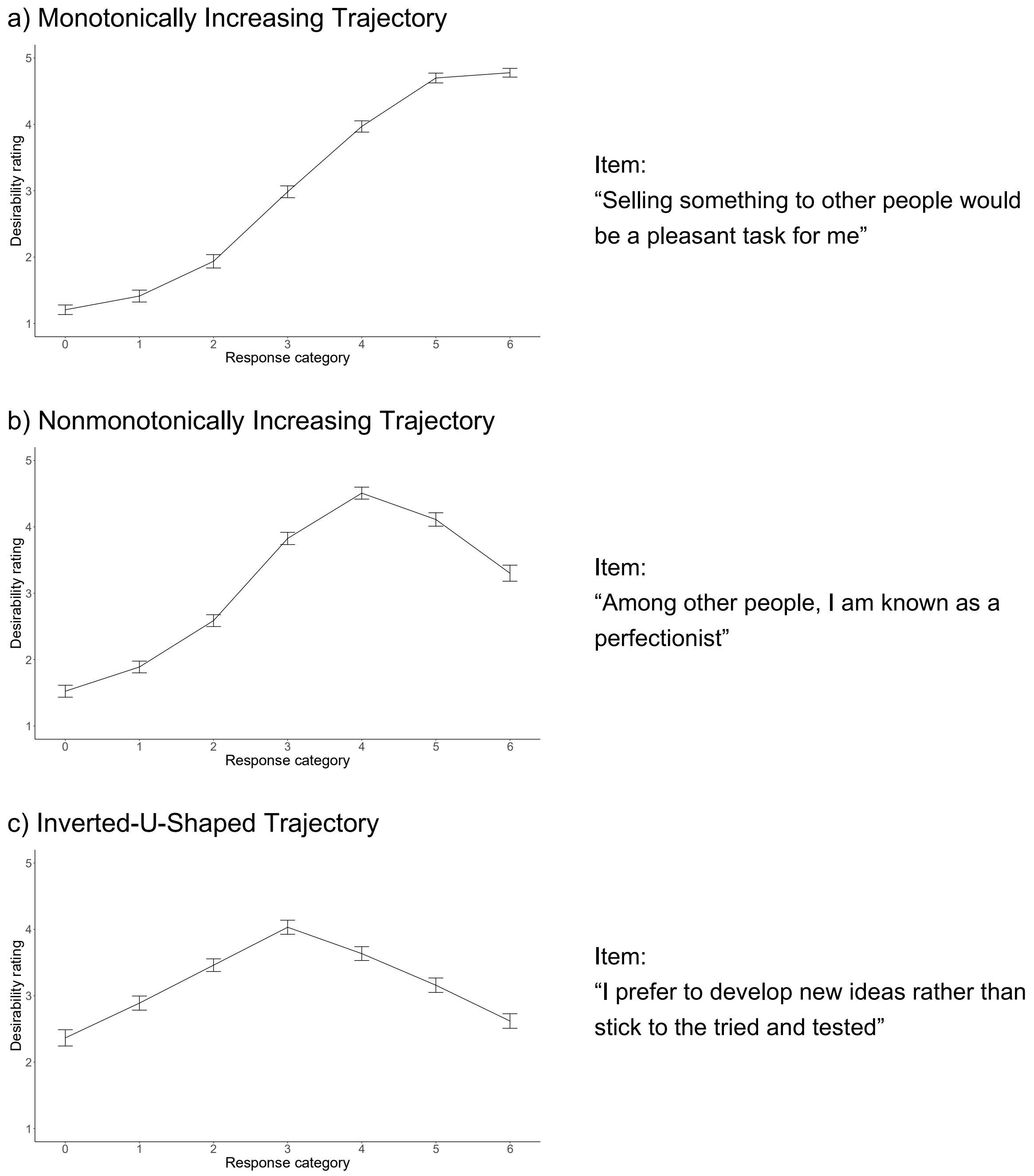
Psychometrically more sophisticated is Böckenholt's (2014) "retrieve-edit-select" (RES) model. This model accounts for motivated misreports in sensitive survey questions, characterized by potential response editing before providing an answer to a question. Technically, the RES model includes for each item a binary latent class variable that indicates whether a test-taker edits his or her retrieved response (i.e., the response based on the substantive trait). In case of a desirable (undesirable) trait being assessed, editing implies that test-takers select response categories that are higher (lower) than the category according to their retrieved response. Whereas the retrieval and editing processes are parameterized through ordinal and binary IRT models, respectively, the selection process is modeled through a matrix of person-specific transition probabilities. A similar model is Leng et al.'s (2020) "retrieve-deceive-transfer" (RDT) model, which has large overlaps with Böckenholt's (2014) RES model but also incorporates the self-deceptive form of SDR in the retrieval process. Both models are theoretically appealing, however, they are computationally demanding and limited to modeling only one substantive trait dimension. Also, they imply a sequential process where faking always comes into play after the retrieval of a trait-based response, which is conceptually questionable.

Recently, Brown and Böckenholt (2022) developed the "faking-as-grade-of-membership" (F-GoM) model. This model should address the limitations of previous latent variable models of faking, especially those of bifactor and mixture modeling approaches as well as of the RES and RDT model. The F-GoM model treats item responses as mixtures of honest and faked responses. Honest ("real") responses are assumed to be influenced by substantive traits, whereas faked ("ideal") responses are assumed to be influenced by a faking factor. Similar to the RES and RDT model, a binary IRT model governed by an editing factor and item characteristics models whether a response is honest or faked. The F-GoM model is not limited to a single substantive trait and allows for varying faking behavior over the course of the questionnaire. At the same time, however, its frequentist estimation curbs how many dimensions can practically be modeled. Also, Brown and Böckenholt (2022) implemented the model in an SEM framework with continuous indicator variables and empirically demonstrated it only using total scale scores as indicators. An implementation and empirical demonstration for individual items is yet to be done.

Despite all methodological advances, the psychometric models outlined above have a common significant limitation: They assume that faking is linearly or at least monotonically related to items. That is, faking is assumed to make higher response categories more likely when an item measures a generally desirable trait and make lower categories more likely when an item measures a generally undesirable trait. However, there is pertinent evidence that social desirability does not for all items increase or decrease monotonically with higher rating scale categories. Kuncel and Tellegen (2009), for instance, instructed participants to rate the social desirability of each response category of each item referring to a specific social context. The finding was that there were many items that measured a desirable trait but did not exhibit strictly monotonically increasing relationships between categories and desirability. Instead, some items showed nonmonotonic (i.e., increasing trend with a decay at the extreme) or even inverted-U-shaped (i.e., highest desirability at the scale midpoint) desirability trajectories. Also, the trajectories varied considerably depending on the social context to which the desirability ratings referred. Similar effects were found by Borkenau et al. (2009), Dunlop et al. (2012), and Seitz, Spengler, and Meiser (2025). An illustration is displayed in Figure 1, which shows the relationships between response categories and desirability ratings for three exemplary items in the context of working as an apprentice in a bank (Seitz, Spengler, & Meiser, 2025). When these item- and category-specific desirability characteristics are ignored, as is the case in the latent variable models of faking outlined above, the faking model is obviously misspecified and does not capture the full process underlying the response bias.

Figure 1

Desirability Trajectories of Three Exemplary Items



a) Monotonically Increasing Trajectory

Item:
“Selling something to other people would be a pleasant task for me”

b) Nonmonotonically Increasing Trajectory

Item:
“Among other people, I am known as a perfectionist”

c) Inverted-U-Shaped Trajectory

Item:
“I prefer to develop new ideas rather than stick to the tried and tested”

Note. Mean desirability ratings are based on $N = 63$ participants. Error bars represent the standard error of the mean. The social context for desirability ratings was an application for a bank apprenticeship. Adapted from "'What If Applicants Fake Their Responses?': Modeling Faking and Response Styles in High-Stakes Assessments Using the Multidimensional Nominal Response Model", by T. Seitz, M. Spengler, and T. Meiser, 2025, *Educational and Psychological Measurement, 85*(4), p. 764 (https://doi.org/10.1177/00131644241307560). CC BY 4.0.

1.2.2 The Multidimensional Nominal Response Model (MNRM) of Faking

To account for such idiosyncratic desirability characteristics, one can make use of the multidimensional nominal response model (MNRM; see Seitz, Spengler, & Meiser, 2025, for an illustration of the approach). Originally, the MNRM was introduced by Takane and de Leeuw (1987) as a multivariate extension of Bock's (1972) univariate approach to modeling nominal (i.e., categorical) data. In the MNRM, item responses are assumed to be influenced by D latent variables. The probability that person n on item i selects response category k out of a set of $K + 1$ categories is modeled as (Falk & Cai, 2016; Thissen & Cai, 2016):

$$p(Y_{ni} = k \mid \boldsymbol{\theta}_n, \boldsymbol{\alpha}_i, \boldsymbol{S}_i, \boldsymbol{\gamma}_i) = \frac{\exp((\boldsymbol{\alpha}_i \circ \boldsymbol{s}_{ik})' \boldsymbol{\theta}_n + \gamma_{ik})}{\sum_{m=0}^{K} \exp((\boldsymbol{\alpha}_i \circ \boldsymbol{s}_{im})' \boldsymbol{\theta}_n + \gamma_{im})} , \qquad (1)$$

$$\text{with } \boldsymbol{\theta}_n = \begin{pmatrix} \theta_{n1} \\ \vdots \\ \theta_{nd} \\ \vdots \\ \theta_{nD} \end{pmatrix}, \ \boldsymbol{\alpha}_i = \begin{pmatrix} \alpha_{i1} \\ \vdots \\ \alpha_{id} \\ \vdots \\ \alpha_{iD} \end{pmatrix},$$

$$\boldsymbol{S}_i = \begin{pmatrix} s_{i10} & \cdots & s_{i1k} & \cdots & s_{i1K} \\ \vdots & & \vdots & & \vdots \\ s_{id0} & \cdots & s_{idk} & \cdots & s_{idK} \\ \vdots & & \vdots & & \vdots \\ s_{iD0} & \cdots & s_{iDk} & \cdots & s_{iDK} \end{pmatrix}, \text{ and } \boldsymbol{\gamma}_i = (\gamma_{i0} \quad \cdots \quad \gamma_{ik} \quad \cdots \quad \gamma_{iK}) .$$

$Y \in \{0, 1, \ldots, k, \ldots, K\}$ denotes a discrete random variable that reflects an observed item response, with k representing its realization (i.e., the selected response category). The specific category probabilities for person n on item i depend on a D-dimensional vector of person parameters ($\boldsymbol{\theta}_n$), a D-dimensional vector of item slopes ($\boldsymbol{\alpha}_i$), a (D×(K+1))-dimensional matrix of scoring weights ($\boldsymbol{S}_i$), and a (K+1)-dimensional vector of item-category intercepts ($\boldsymbol{\gamma}_i$). Person parameters θ_{nd} are the score of a person n on latent dimension d. Item-category intercepts γ_{ik} reflect the propensity toward category k on item i for a person with person parameters of 0. Item slopes α_{id} represent the relationship between item i and dimension d. Scoring weights s_{idk}, in contrast, represent the relationship between category k and dimension d on item i. Vector $\boldsymbol{\alpha}_i$ and column vector $\boldsymbol{s}_{ik}$ from matrix $\boldsymbol{S}_i$ are connected through the Hadamard product (denoted by ∘), such that parameters that pertain to the same dimension are multiplied. The resulting vector is transposed and multiplied by vector $\boldsymbol{\theta}_n$, leading to a sum of products $\alpha_{id}\, s_{idk}\, \theta_{nd}$ over the D dimensions. After adding γ_{ik} to this sum, the resulting term is transformed through a multinomial logistic function (aka softmax function) to a range from 0 to 1, which yields the model-implied probability of an item response. Thus, the MNRM belongs to the class of divide-by-total IRT models (Thissen & Steinberg, 1986). Table 1 in Seitz, Spengler, and Meiser (2025) provides an overview of the MNRM's parameters.

The D latent dimensions are usually assumed to follow a multivariate normal distribution with expectation vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. However, estimating all model parameters freely is not possible because the model is not identified without any constraints (see Falk & Cai, 2016; Henninger & Meiser, 2020; Johnson & Bolt, 2010, for details). Typical identification constraints are to fix the latent dimensions' expectations to 0 and their variances to 1. Also, one intercept per item has to be fixed for identification, which is typically achieved by fixing the intercept of the first category to 0 for all items.

Furthermore, scoring weights can be set to appropriate values in order to specify the meaning of the latent dimensions that are modeled. Because scoring weights reflect how categories are related to dimensions, theoretical considerations can guide the specification of scoring weights. Following the Likert scale logic, where higher categories are associated with higher trait levels, scoring weights of a dimension representing a substantive trait are typically set to equally-spaced values. For a 7-point Likert scale, a scoring weight vector of $(0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$ can be specified for every item measuring the respective substantive trait. If only one substantive trait and no additional dimensions are modeled, such a specification of scoring weights yields a model equivalent to a generalized partial credit model (GPCM; Muraki, 1992) or, if item slopes are constrained to be equal across items, a partial credit model (PCM; Masters, 1982). When modeling response styles, scoring weights can be set to 1 and 0, depending on whether a category is triggered by the particular response style or not (see Falk & Cai, 2016; Henninger & Meiser, 2020). For example, $(1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1)$ can be specified for an ERS dimension, $(0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0)$ for an MRS dimension, and $(0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1)$ for an ARS dimension. Since response styles reflect stable preferences for certain types of response categories irrespective of the item content, the same scoring weight vector per response style dimension is usually specified for all items of the questionnaire.

Analogously, one can specify scoring weights for a faking dimension (Falk & Cai, 2016; Seitz, Spengler, & Meiser, 2025). The specification is, however, not as straightforward as for substantive traits and response styles. Conceptually, scoring weights of faking should represent the item-specific desirability levels of the different response categories with respect to the social context in which the assessment takes place. As elaborated above, desirability trajectories vary between items to the extent that higher categories are not related to higher desirability levels for all items (e.g., Kuncel & Tellegen, 2009; see Figure 1). Scoring weight vectors of faking are hence neither constant across items nor globally redundant to scoring weight vectors of substantive trait dimensions. The following matrix contains the scoring weights of item *i*, which features a 7-point Likert scale and measures the first out of five substantive traits while ERS and faking are accounted for:

$$\boldsymbol{S}_i = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ des_{i0} & des_{i1} & des_{i2} & des_{i3} & des_{i4} & des_{i5} & des_{i6} \end{pmatrix}. \qquad (2)$$

To set scoring weights of faking (des_{ik}) in an empirical setting, one can utilize data from pilot studies. For instance, des_{ik} can be the mean desirability rating from a pilot study where participants rate each category of each item with respect to desirability. This is what Seitz, Spengler, and Meiser (2025) did, and has also been done in Articles II and III of this dissertation. Alternatively, when conducting a pilot study where participants judge for each item which category is most desirable, des_{ik} can be the relative frequency with which a category is judged as most desirable on a given item. This is what has been done in Article I.

Modeling faking by means of the MNRM has several appealing features. From a psychometric perspective, it improves the measurement of substantive traits by accounting for faking based on the item-, category-, and context-specific nature of social desirability.[2] In contrast, previous latent variable models make overly simplified assumptions about how faking manifests in item responses: SEM approaches like the bifactor model or F-GoM model assume a linear relationship between a faking factor and all items; the RES and RDT model assume that faking always increases (decreases) responses when a desirable (undesirable) trait is measured; the aforementioned IRTree approaches equate faking with extreme responding; and exploratory mixture models only yield discrete faking classes if the post-hoc interpretation of the observed latent classes gives rise to this (cf. Section 1.2.1). Consequently, the MNRM approach should provide estimates of substantive trait scores that are more appropriately

[2] Note that the MNRM represents a genuine dominance IRT model as opposed to an ideal-point IRT model. Ideal-point models assume that item endorsement increases the closer a person's trait level is to an item's location on the trait continuum (e.g., Chernyshenko et al., 2007; Roberts et al., 2000), such that intermediate as opposed to extreme trait levels can have the highest expected item response. This is reflected in nonmonotonic item response functions (IRF). IRFs are, however, not to be confused with item desirability trajectories, which depict how response categories are related to social desirability at a specific item. Because the MNRM is a dominance model, higher substantive trait scores always imply higher probabilities of selecting higher categories, while higher faking scores always imply higher probabilities of selecting more desirable categories – though the most desirable categories may not necessarily be the highest categories on the rating scale.

adjusted for faking. In applied measurement contexts, such as high-stakes testings, an appropriate adjustment of substantive trait scores does not only help to increase fairness, but also helps to ensure that decision-makers can base their decisions on scores that better capture the constructs intended to be measured. Moreover, the MNRM approach is valuable from a substantive research perspective. As opposed to approaches aiming at faking prevention, faking is not eliminated from the outset but disentangled from substantive traits on a latent level, which allows for investigations of the substantive nature of the faking construct. For instance, latent correlations of faking with substantive traits or other dimensions can be estimated to answer the question of whether faking represents a mere nuisance variable or a construct with psychological meaning that can be integrated into the nomological network of interindividual difference variables.

1.3 Open Questions and Necessary Model Extensions

In my master thesis, which has eventually been published in Seitz, Spengler, and Meiser (2025), I conducted an initial investigation of the MNRM approach to modeling faking. This investigation featured an empirical application of the model in a high-stakes dataset of $N = 3046$ job applicants who had taken a Big Five personality test as part of their application for an apprenticeship at a large German bank. In this empirical analysis, modeling faking significantly increased model fit over and above response styles, and improved the personality scales' discriminant validity by disinflating the estimated correlations between the substantive traits. An application of the model to a low-stakes dataset of job incumbents also provided initial validation evidence that the model indeed captures faking and adjusts substantive trait scores in the expected direction. In addition, the faking dimension exhibited significant relationships with covariates, including agreeableness and intelligence.

Notwithstanding the promising initial results, several open questions and necessary model extensions have remained, which form the foundation of the three research projects of this dissertation: First, it has remained unclear under which conditions the MNRM can effectively adjust model parameters for faking, especially with regard to the way item content needs to be related to social desirability for the model to perform well (Article I; Seitz et al., 2024). Second, the MNRM as described above assumes the same measurement model for all test-takers. Some test-takers, however, might not fake at all while others do not at all consider substantive traits, which could bias parameters and diagnostic inferences unless the model accounts for such different response strategies (Article II; Seitz, Alagöz, & Meiser, 2025). Third, the employed response strategy of a test-taker might not be constant over the course of the questionnaire. A model that also allows for dynamic switches between strategies, however, could require additional data to yield satisfactory results (Article III; Seitz & Ulitzsch, 2025). In the following sections, I will in more detail delineate the motivation, technical aspects, methods, results, and limitations of each of the three projects. The final section will then discuss and integrate the findings, elaborate on psychometric implications for research and measurement practice, and derive directions for future research.

2 Article I

Seitz, T., Wetzel, E., Hilbig, B. E., & Meiser, T. (2024). Using the multidimensional nominal response model to model faking in questionnaire data: The importance of item desirability characteristics. *Behavior Research Methods, 56*(8), 8869–8896. https://doi.org/10.3758/s13428-024-02509-x

In the context of psychometric modeling, it is essential to show that a model's parameter estimates are unbiased (i.e., are not systemically higher or lower than the true parameter values) and more accurate (i.e., have smaller error) than the estimates of competitor models. Such a question can only be answered based on a simulation study, where the true parameter values are known. However, as mentioned above, the initial investigation of the MNRM approach to modeling faking from my master thesis was focused on an empirical application of the model and provided only quasi-experimental validation evidence.[3] Hence, essential psychometric properties have remained unknown, especially the question of whether modeling faking with the MNRM can indeed afford more accurate representations of person parameters and latent correlations compared to other models.

To address these open questions, we conducted a simulation study in Article I. In this simulation, we investigated factors that facilitate or limit the performance of a model accounting for faking compared to models not accounting for faking. The simulation included the factors test length, sample size, presence of response styles, and impact of faking. Of particular interest, however, has been another factor that might have a systematic effect on how well faking can be modeled using the MNRM, namely the role of variation in desirability characteristics across items. Item desirability characteristics refer to the way item content is related to social desirability, that is, the trajectory of desirability across response categories (see Figure 1). Although Kuncel and Tellegen (2009) found that items of personality tests differ in terms of their desirability trajectories, the typical case is that higher categories are associated with higher desirability levels. That is, personality items are in most cases constructed such that descriptive aspects of the substantive traits are confounded with evaluative aspects (Peabody, 1967), which implies that the selection of high response categories can stem from high substantive trait levels, high faking levels, or both. Transferred to modeling faking using the MNRM, a situation with confounded descriptive and evaluative aspects causes high collinearity between the scoring

[3] Note that the preprint (Seitz et al., 2023) and not the final publication (Seitz, Spengler, & Meiser, 2025) of this initial investigation of the model is cited in Article I.

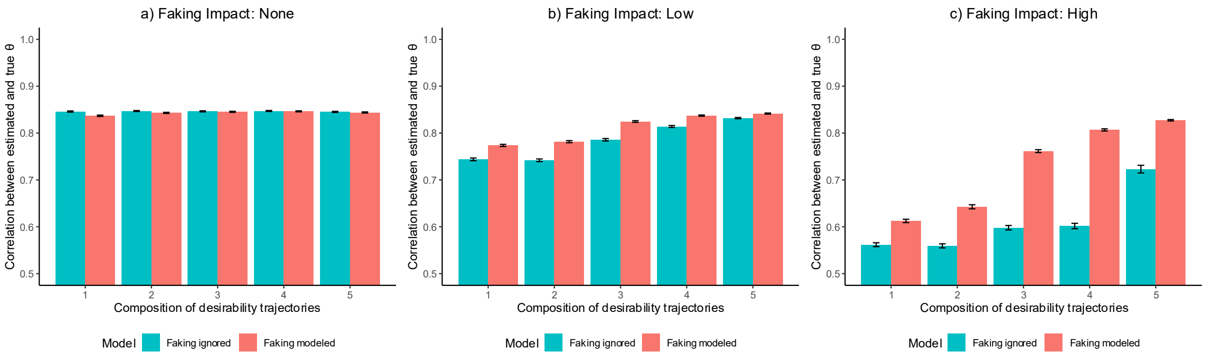
weight vectors of substantive traits and faking. Considering this collinearity, it should be increasingly difficult to disentangle substantive traits and faking the more items there are with highly overlapping scoring weight vectors. In contrast, if descriptive and evaluative aspects were not associated across items, the model-based separation should be facilitated, and high item responses might be a better indication of high substantive trait levels even when faking is not statistically accounted for.

The main results regarding the recovery of person parameters from the simulation study in Article I are displayed in Figure 2 (analogous results regarding the recovery of latent correlations). When faking is part of the data-generating process, the simulation results indicate that parameters can overall be better recovered by a model with faking dimension compared to a model without faking dimension. The extent to which a model with faking dimension improves recovery, however, strongly depends on the impact of faking in the data, as simulation conditions with a low faking impact yielded only small differences in parameter recovery between models while conditions with a high faking impact yielded much larger differences. Crucially, the results also show that a model including a faking dimension is not inferior to the correctly specified population model when there is truly no faking in the data. The faking dimension in the MNRM hence does not seem to erroneously absorb substantive variance in the sense that the estimation of substantive trait scores is impaired when the model is actually overparameterized. In contrast, this constitutes a major drawback of the popular SDR scales (e.g., de Vries et al., 2014; McCrae & Costa, 1983) and has similarly been found in the context of response styles (Merhof et al., 2024).

Concerning item desirability characteristics, the simulation results suggest that item desirability characteristics moderate the effect of modeling faking (see Figure 2), whereas there seem to be no interactions with test length, sample size, and the presence of response styles in the data. In case of a low faking impact, the superiority of the model with faking dimension was most pronounced when all items had monotonically increasing desirability trajectories (item composition 1 in Figure 2), and almost vanished when there was large variety in desirability trajectories across items (item composition 5 in Figure 2). In case of a high faking impact, the model with faking dimension was superior in all item compositions. Note, however, that item desirability characteristics also had a pronounced main effect. As item compositions with more variety in desirability trajectories were associated with an overall better parameter recovery, the results indicate that reducing the overlap between scoring weight vectors of substantive traits and faking does not only facilitate the disentanglement of the latent dimensions in a psychometric model of faking, but can also improve the measurement of substantive traits when faking is not statistically modeled. According to the simulation, modeling faking will nevertheless be beneficial if the impact of faking in the data is high.

Figure 2

Recovery of Substantive Trait Scores in the Simulation Study From Article I



Note. The depicted recovery of substantive trait scores applies to the representative case of 6 items per substantive trait scale, a sample size of 1000, and extreme response style (ERS) being present in the data. The five compositions of desirability trajectories denote five item compositions with increasing variety in desirability trajectories across items. Models ignoring faking only included dimensions for substantive traits and ERS, whereas models accounting for faking also included a faking dimension. Values reflect the back-transformed mean of the Fisher-*z*-transformed correlations between estimated and true person parameters across replications within a condition. Error bars represent the standard error of the mean. Reprinted from "Using the Multidimensional Nominal Response Model to Model Faking in Questionnaire Data: The Importance of Item Desirability Characteristics", by T. Seitz, E. Wetzel, B. E. Hilbig, and T. Meiser, 2024, *Behavior Research Methods, 56*(8), p. 8876 (https://doi.org/10.3758/s13428-024-02509-x). CC BY 4.0.

In Article I, we also examined whether the effects of item desirability characteristics can similarly be found in empirical questionnaire data. Therefore, we adapted items from the widely-used *Big Five Inventory 2* (BFI-2; Danner et al., 2016, 2019) in a way that they should still measure the Big Five but deconfound descriptive and evaluative aspects. We therefore created more items with nonmonotonically increasing, inverted-U-shaped, and even decreasing desirability trajectories. After piloting both the original and new items, classifying them into the different types of desirability trajectories, and composing five item sets with different variety in desirability trajectories, we collected data from $N = 1070$ participants who responded to all items under both a low-stakes (LS) and experimental high-stakes (HS) condition. In the LS condition, participants should respond as honestly as possible. The HS condition, in contrast, featured a hypothetical application scenario in which participants should respond to the items as if they were applying for a leadership position in the industry. To create actual stakes for participants and to approximate the circumstances of a real-life job application, we offered a financial incentive for faking and used tailored instructions in the HS condition.

In all considered item sets, the model including a faking dimension yielded a significantly higher model fit than a model accounting only for substantive traits and a model accounting for substantive traits and response styles. Additionally, because data from the same persons under both an LS and HS condition was available, the question of whether modeling faking indeed improves the measurement of substantive traits could be investigated empirically. The only necessary assumption for such an investigation is that estimates of substantive trait scores in an LS condition are not systematically biased by faking. Under this assumption, higher LS-HS correlations indicate that one model's HS substantive trait score estimates are a better representation of the actual substantive trait levels than another model's estimates. Indeed, 22 out of the 25 considered LS-HS correlations were descriptively higher when the HS substantive trait scores were estimated using a model with faking dimension than when they were estimated using a model without faking dimension. 19 of these correlation differences reached statistical significance. This is evidence that the faking dimension in the MNRM effectively adjusts biases in the rank order of test-takers.[4] At the same time, the effect of modeling faking interacted with item desirability characteristics like in the simulation conditions with a low faking impact (Figure 2b). In particular, the effect was most pronounced when items predominantly had increasing desirability trajectories, while more variety in desirability trajectories across items led to generally higher LS-HS correlations.

To conclude, the findings from Article I indicate that modeling faking is worthwhile, as it increasingly improves parameter recovery the higher the faking impact is, and does not worsen recovery when faking is completely absent in the data. The effect of modeling faking interacts with item desirability characteristics, which are themselves associated with different levels of parameter recovery. This suggests that developing tests that contain items with varying desirability characteristics can remedy the negative psychometric effects of faking already at the stage of item construction, even if faking is not statistically accounted for. However, potential pitfalls shall be considered as well. On the one hand, the MNRM approach to modeling faking comes with general limitations, which I will elaborate on in Section 6.2. On the other hand, creating more variety in item desirability trajectories can be a challenging task when the to-be-measured trait is inherently desirable or undesirable in a given context. Wood et al. (2022, p. 818) noted "that the social desirability of personality traits is partially intrinsic and partially the result of item writing practices". That is, for inherently desirable (undesirable) personality traits, it will be hard or even impossible to create items with inverted-U-shaped and/or

[4] I have conducted an additional analysis of this kind that compares different models of faking concerning the empirical recovery of LS estimates of substantive trait scores. Details and results are reported in Section 5.

decreasing (increasing) desirability trajectories that do not change the meaning of the construct. It is hence crucial to review and potentially adapt such newly created items to ensure that they are in line with the construct definition of the respective substantive trait. In addition, it depends on the actual goal how much subtle change in the meaning of the construct can be accepted. If the goal is to measure narrowly defined personality traits in research contexts, the proposed item refinements will be less appropriate than in applied measurement contexts like personnel selection, where the goal is to have fair assessments that are not contaminated by faking.

3 Article II

Seitz, T., Alagöz, Ö. E. C., & Meiser, T. (2025). Disentangling qualitatively different faking strategies in high-stakes personality assessments: A mixture extension of the multidimensional nominal response model. *Educational and Psychological Measurement, 85*(6), 1237–1277. https://doi.org/10.1177/00131644251341843

The MNRM of faking as introduced in Section 1.2.2 and investigated in Article I treats faking as a continuous latent variable that quantifies each test-taker's faking degree. This is a feature the MNRM shares with other faking models, such as the bifactor model, as well as models of response styles (see Falk & Cai, 2016). Put differently, the MNRM assumes that all test-takers use the same qualitative response strategy but vary quantitatively in the degree of faking. However, there are several findings in the literature suggesting that test-takers in high-stakes assessments in fact use qualitatively different response strategies. Most studies of faking prevalence, for instance, have found that many test-takers do engage in faking whereas a considerable proportion do not show self-presentational behavior (e.g., Donovan et al., 2003; Griffith et al., 2007). Relatedly, studies examining the thought process in high-stakes assessments have revealed that test-takers differ in what guides their responses (e.g., Robie et al., 2007). It is hence questionable if the response process associated with faking can be described by a single measurement model with a continuous faking dimension. Consider, for example, a test-taker whose responses are only a function of substantive trait levels (i.e., who does not engage in faking). The MNRM will for this test-taker estimate substantive trait scores that are adjusted for an estimated faking level even though this person has not faked at all. Likewise, a test-taker whose responses are only a function of his or her faking level will receive substantive trait score estimates even though the person has not at all considered substantive traits in the responses.

In Article II, a mixture MNRM (M-MNRM) is proposed that allows to disentangle qualitatively different faking-related response strategies. The M-MNRM constitutes a confirmatory mixture model with theory-driven constraints on class-specific parameters (cf. Alagöz & Meiser, 2024). In particular, the model includes the following three latent classes that represent the response strategies test-takers can use in high-stakes assessments (cf. Robie et al., 2007): an "S-only class" (responses as a function of substantive traits only), an "S&F class" (responses as a function of substantive traits and faking), as well as an "F-only class" (responses as a function of faking only). The full model can be denoted as:

$$p(Y_{ni} = k \mid \boldsymbol{\theta}_n, \boldsymbol{\alpha}_{ic}, \boldsymbol{S}_i, \gamma_{ic}) = \sum_{c=0}^{2} \frac{\exp((\boldsymbol{\alpha}_{ic} \circ \boldsymbol{s}_{ik})'\boldsymbol{\theta}_n + \gamma_{ikc})}{\sum_{m=0}^{K} \exp((\boldsymbol{\alpha}_{ic} \circ \boldsymbol{s}_{im})'\boldsymbol{\theta}_n + \gamma_{imc})} p(\zeta_n = c) \ , \tag{3}$$

with ζ_n reflecting the class membership of person n. This equation describes the total probability of response k for person n on item i. The term $p(\zeta_n = c)$ represents the unconditional membership probability of class c, which reflects the proportion of class c across all test-takers. Parameters with index c are, in principle, class-specific. However, since the model entails a confirmatory modeling of latent classes, particular parameter constraints are imposed. Namely, to implement the three classes based on their definitions, slopes of latent dimensions that are not part of a given response strategy are set to 0 for the respective class. Also, in order for the latent dimensions to have the same meaning across classes, non-fixed item slopes are constrained to be class-invariant. The M-MNRM furthermore allows for modeling relationships between class membership and external covariates, which is parameterized through a latent multinomial logistic regression. Model estimation can be performed with a Bayesian Markov chain Monte Carlo (MCMC) procedure using the software *JAGS* (Plummer, 2017) via the R package *runjags* (Denwood, 2016).

In a simulation study with different class proportion conditions, we investigated the M-MNRM regarding parameter recovery, its superiority over non-mixture models, as well as the performance of Bayesian model selection criteria. The results of the simulation show good parameter recovery of the M-MNRM in conditions with multiple latent classes, whereas non-mixture models yield biased and less accurate parameter estimates. In addition, the M-MNRM can afford high classification accuracy, irrespective of the actual class proportions. Importantly, however, the results also indicate that the M-MNRM is not inferior to the correctly specified non-mixture model when the data generation features only a single class. Considering the accurate estimation of class proportions and an acceptable performance of model selection criteria in the simulation, researchers and practitioners thus have a good basis to decide if the mixture model extension is justified in a particular dataset.

We also applied the M-MNRM to three empirical datasets ($1824 \leq N \leq 3046$) from different personnel selection contexts. In all datasets, the M-MNRM was selected over the different non-mixture models, yielded a better description of the data according to posterior predictive model checks (PPMC), and exhibited high class separability according to model entropy. Across datasets, the M-MNRM estimated that approximately 50% of test-takers were members of the “S&F class”, whereas about 40% were “S-only class” and about 10% were “F-only class” members. Because the “S&F class” has the same measurement model as the non-mixture MNRM, these class proportion estimates suggest that the non-mixture version of the MNRM is misspecified for roughly 50% of test-takers in high-stakes assessments. Class membership additionally turned out to be consistently associated with the covariates of integrity and intelligence, with test-takers in the “S-only class” having the highest integrity and intelligence scores and test-takers in the “F-only class” having the lowest scores on integrity and intelligence. Also, validation analyses provided evidence for the plausibility of the class assignments performed by the M-MNRM. In particular, response distributions differed plausibly between classes (e.g., higher mean responses in the “F-only class” than in the other classes for items with monotonically increasing desirability trajectories), and estimated class proportions in a low-stakes sample from a career counseling context were expectedly different (e.g., higher “S-only class” proportion) compared to the high-stakes datasets from the personnel selection contexts.

In summary, the proposed mixture model in Article II constitutes a valuable extension of the MNRM approach to modeling faking. It is not limited to a single measurement model and thus allows for modeling faking in terms of both a continuous and discrete variable. This aligns with Kiefer and Benit’s (2016) conclusion that a combined use of quantitative and qualitative modeling techniques would describe the current understanding of faking best. In applied measurement contexts, this combined use allows, on the one hand, for a more appropriate estimation of substantive trait scores, because the estimation is based on a measurement model that is in line with the response behavior of a test-taker. On the other hand, it allows for an individual strategy classification of the responses given by a particular test-taker, which can be a valuable piece of information regarding the trustworthiness of this person’s responses. Nonetheless, some limitations of the M-MNRM should be noted: First, direct comparability of person parameters across classes might be limited because of the non-invariance of item-category intercepts. As is the case for all latent class models, parameters must be on a common scale to be meaningfully comparable across classes (Paek & Cho, 2015). To establish a common scale of person parameters despite the non-invariance of intercepts, there are different options. One option is to model a set of anchor items with truly invariant item parameters across classes (e.g., items with neutral social desirability). A common scale can also be achieved if the classes have the same true latent means. In this case, a class-invariant identification of the origin of the latent variables in the M-MNRM is in line with the truth, such that person parameters should be on the same scale and hence be comparable across classes. A second limitation of the M-MNRM as proposed in Article II is that, because of its confirmatory definition of the three latent classes, more fine-grained differences in faking behavior might not yet be satisfactorily captured. Röhner et al. (2025), for instance, identified as many as 35 distinct faking strategies. Such a high number of latent classes will be difficult or even impossible to implement in a mixture model, however, future studies could apply mixture modeling of high-stakes assessment data less restrictively to examine heterogeneity in a more exploratory way.

3.1 Additional Analysis: Experimental Validation of Class Assignments

As mentioned above, Article II provides initial validation evidence regarding the assignment of test-takers to the three latent classes. Going beyond Article II, I have run an additional analysis to validate the M-MNRM's class assignments experimentally. I have therefore made use of the dataset from Article I, which includes item responses under an LS (honest condition) and HS condition (hypothetical application condition), and expected a higher estimated "S-only class" proportion and lower estimated "S&F class" and "F-only class" proportions in the LS than in the HS condition. Because of the experimental nature of this dataset, such differences in estimated class proportions between conditions can be causally attributed to the different stakes between conditions. In contrast, a quasi-experimental comparison like in Article II holds the risk that such differences in class proportion estimates between conditions are due to confounds, such as different sample characteristics.

For the analysis, I fitted the M-MNRM separately to the item responses from the LS and HS condition using the estimation procedure described in Article II, and then compared the estimated class proportions.[5] Table 1 displays the results. As expected, the estimated "S-only class" proportion was higher in the LS condition than in the HS condition, whereas the estimated "S&F class" proportion was lower in the LS condition than in the HS condition. These differences were statistically meaningful as credible intervals did not overlap. The estimated "F-only class" proportion was descriptively also lower in the LS condition than in the HS condition. Thus, this analysis adds internally valid evidence regarding the response strategy classifications performed by the M-MRNM.

[5] To avoid computational overload, I did not consider all available 125 items per condition, but used the 60 items selected in a bachelor thesis under my supervision (Geyer, 2023). This bachelor thesis had the goal of composing an item set for a refined Big Five test that, on the one hand, yields satisfactory item and test statistics and, on the other hand, features variety in desirability trajectories across items.

Table 1

Class Proportions in the Experimental Data From Article I Estimated by the Model From Article II

	LS condition	HS condition
“S-only class”	79.2% [76.3%, 81.9%]	39.8% [35.8%, 44.0%]
“S&F class”	18.4% [15.8%, 21.2%]	57.1% [52.8%, 61.3%]
“F-only class”	2.4% [1.6%, 3.5%]	3.1% [2.0%, 4.3%]

Note. N = 1070. Values in brackets represent the 95% credible interval. LS = low-stakes; HS = high-stakes.

4 Article III

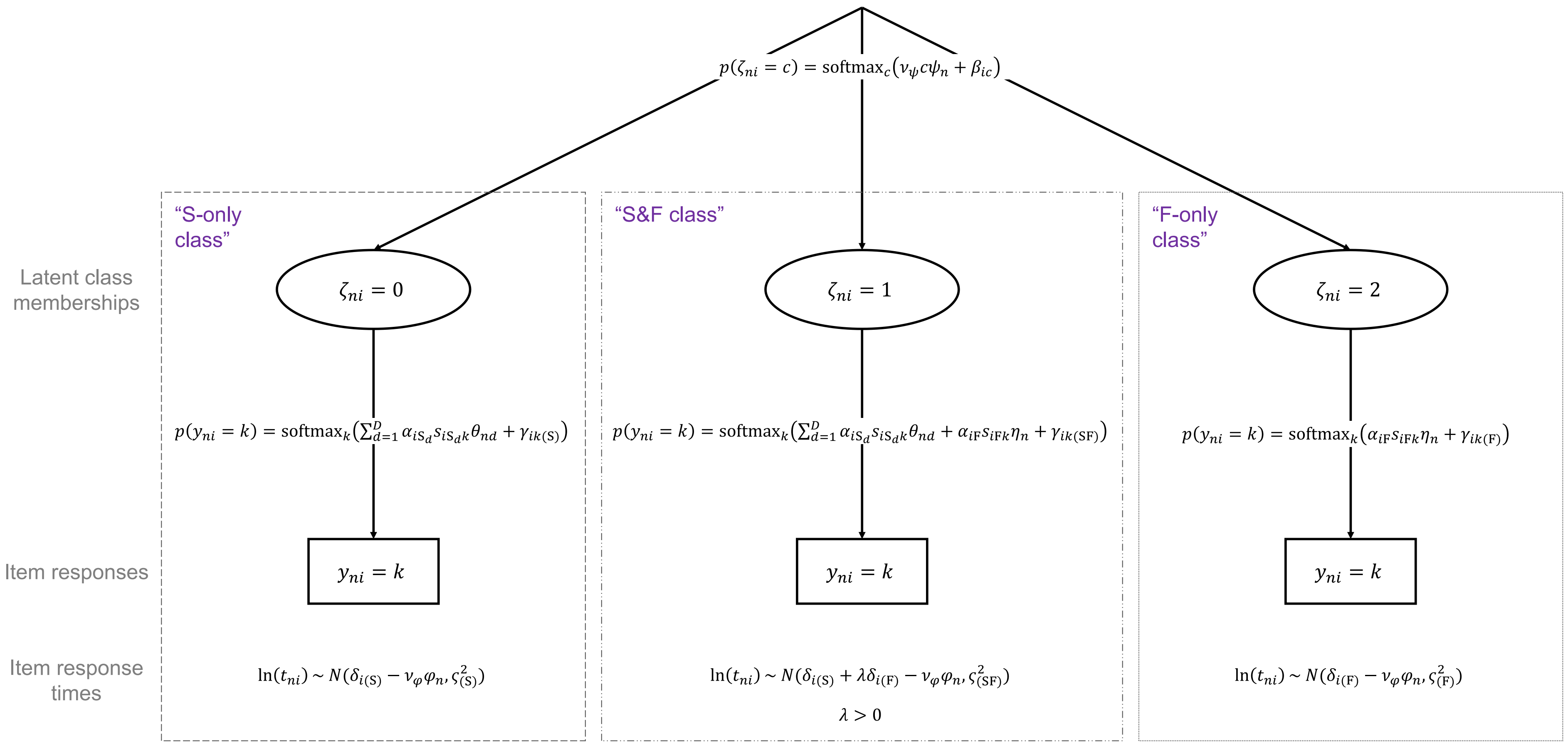
Seitz, T., & Ulitzsch, E. (2025). *Faking in high-stakes personality assessments: A response-time-based latent response mixture modeling approach* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

As described above, most faking models quantify the degree of faking on a latent continuum, others assign test-takers to latent classes representing qualitatively different faking-related response behaviors. The mixture extension of the MNRM in Article II combines quantitative and qualitative modeling techniques of faking, but still treats faking as a person variable that is constant throughout the entire test. There are, however, several reasons why the stability of faking across test items can be questioned: First, faking is a complex interaction of person and situation characteristics like ability, opportunity, and motivation to fake (Tett & Simonet, 2011), such that actual faking behavior in personnel selection is likely to vary between items when some items are seen as more instrumental than others for attaining the goal of making a good impression (e.g., Ellingson & McFarland, 2011). Second, studies in the context of lying research have shown that people behave dishonestly to the extent that they profit, but not to the extent that they have to give up their self-concept of being an honorable person (Mazar et al., 2008). Third, conflicting motives of test-taking behavior (e.g., convincing a prospective employer of one's suitability for the job vs. staying true to oneself; Kuncel et al., 2011) can lead test-takers to fake at some items but be honest at other items.

Article III presents a psychometric model that accounts for heterogeneity in faking behavior between both persons *and* items (see Figure 3 for the full model). The proposed model is again a mixture model with the three latent classes from Article II. However, class membership is not restricted to be constant for a given test-taker, but modeled through an ordinal latent response model based on an additional person parameter (strategy inclination; ψ_n) and item characteristics (item-class intercepts; β_{ic}). At the same time, the proposed model makes use of item-level response times (RT), which are modeled through a person speed parameter (φ_n) and item parameters that reflect how long an item takes to be responded to in a particular class (item time intensity parameters; δ_{ic}). In the face of inconsistent findings in the literature concerning the question of whether faking increases or decreases RTs (e.g., Holden et al., 1992; Holtgraves, 2004), no constraints are put on time intensities in the "S-only class" and "F-only class". However, because both substantive trait responding and faking are involved in the "S&F

Figure 3

The Proposed Response-Time-Based Latent Response Mixture Model of Faking From Article III


$p(\zeta_{ni} = c) = \text{softmax}_c(\nu_\psi c\psi_n + \beta_{ic})$
Latent class memberships
Item responses
Item response times
"S-only class"
$\zeta_{ni} = 0$
$p(y_{ni} = k) = \text{softmax}_k(\sum_{d=1}^{D} \alpha_{i\text{S}_d} s_{i\text{S}_d k}\theta_{nd} + \gamma_{ik(\text{S})})$
$y_{ni} = k$
$\ln(t_{ni}) \sim N(\delta_{i(\text{S})} - \nu_\varphi \varphi_n, \varsigma^2_{(\text{S})})$
"S&F class"
$\zeta_{ni} = 1$
$p(y_{ni} = k) = \text{softmax}_k(\sum_{d=1}^{D} \alpha_{i\text{S}_d} s_{i\text{S}_d k}\theta_{nd} + \alpha_{i\text{F}} s_{i\text{F}k}\eta_n + \gamma_{ik(\text{SF})})$
$y_{ni} = k$
$\ln(t_{ni}) \sim N(\delta_{i(\text{S})} + \lambda\delta_{i(\text{F})} - \nu_\varphi \varphi_n, \varsigma^2_{(\text{SF})})$
$\lambda > 0$
"F-only class"
$\zeta_{ni} = 2$
$p(y_{ni} = k) = \text{softmax}_k(\alpha_{i\text{F}} s_{i\text{F}k}\eta_n + \gamma_{ik(\text{F})})$
$y_{ni} = k$
$\ln(t_{ni}) \sim N(\delta_{i(\text{F})} - \nu_\varphi \varphi_n, \varsigma^2_{(\text{F})})$

Note. Notational elements not introduced in the text: t_{ni} = response time of person n on item i; ν_ψ = item-invariant strategy inclination slope; ν_ψ = item-invariant speed slope; ς_c^2 = residual variance of log-RTs in class c.

class", time intensities of this class are constrained to be a function of "S-only class" and "F-only class" time intensities: $\delta_{i(\mathrm{SF})} = \delta_{i(\mathrm{S})} + \lambda\delta_{i(\mathrm{F})}$. To encode the assumption that RTs in the "S&F class" are on average longer than RTs in the other two classes (Walczyk et al., 2003), the proportionality constant λ is constrained to be positive.[6] Like the model in Article II, this RT-based person-by-item mixture model can be estimated with a Bayesian MCMC algorithm using JAGS.

The proposed model has some overlap with Brown and Böckenholt's (2022) F-GoM model, in which response strategy use per person also does not have to be constant across items. However, the F-GoM model only includes an honest and a faking class, and has yet to be extended to the case of modeling individual item responses and not total scale scores as indicator variables. Our proposed mixture model of faking, in contrast, includes a class where substantive traits *and* faking influence item responses, which is conceivable based on the finding that some test-takers in high-stakes assessments base their responses on both their true personality and the ideal applicant (Robie et al., 2007) as well as on research that has found faking to operate at the editing stage of item responding (Holtgraves, 2004; Walczyk et al., 2003). Also, as opposed to the F-GoM model, the proposed model can explicitly account for nonlinear and nonmonotonic relationships between categories and desirability. An additional feature other faking models do not have is that the proposed model incorporates item-level RTs in the context of high-stakes assessments (see Ulitzsch et al., 2020; Ulitzsch, Pohl, et al., 2022, for similar models accounting for disengaged and careless responding in low-stakes assessment contexts). Including RTs serves two purposes: On the one hand, it should facilitate model estimation and class separation, since the information for individual class assignment (i.e., a single item response) would otherwise be very sparse. On the other hand, the inclusion of RTs allows investigating the response process associated with faking in a sophisticated way. This can help to address substantive research questions, such as the question of whether faking actually increases or decreases RTs. In addition, the model as a whole allows examining further substantive issues, including the question of how different faking tendencies covary with substantive person characteristics and the question of what items are particularly prone to different response strategies.

Using a simulation, we examined the proposed model concerning its ability to accurately recover parameters and compared it to less complex models. To make the simulation representative of actual high-stakes data conditions, we used parameter estimates from this

[6] The assumption of longest RTs in the "S&F class" rests on theoretical considerations as well as on an experimental pre-study we conducted that corroborated this assumption.

article’s empirical demonstration as data-generating parameter values in the simulation. Overall, the simulation results indicate good parameter recovery of the proposed model in terms of negligible bias and decent estimation accuracy. Models that lack components of the data-generating process, in contrast, produce systemically biased and much less accurate parameter estimates. Mixture models with constant class membership per person, for example, strongly overestimate the proportion of the “S&F class”. Also, modeling RTs indeed seems to be advantageous for model convergence as well as parameter recovery. Nevertheless, considering the size of the observed hit rates in the simulation, the results indicate that response strategy classifications of individual item responses should be interpreted with caution when strategy use truly varies within persons. In additional simulations reported in the Supplement of Article III, we fitted the different models to simulated datasets in which strategy use was truly constant per person or truly constant across the entire sample. The results of these additional simulations indicate that classification accuracy is generally much higher in populations in which strategy use does not vary between items. However, although the proposed latent response model is flexible enough to approximate constant class membership per person and even for the whole sample, the additional simulations also suggest that the model nevertheless yields slightly less accurate parameter estimates compared to non-overparametrized models in populations with constant strategy use across items. That is, the proposed model seems to be prone to overfitting in such conditions.

Article III also includes an empirical demonstration of the proposed model with data from *N* = 1824 applicants for a police officer traineeship at a German police department. In line with the simulation results, not modeling RTs in a mixture model where class membership could vary within persons led to convergence problems, whereas no such issues occurred in the proposed model with RTs. Compared to different non-mixture models and the mixture model with constant class membership per person, the proposed model yielded the best compromise between fit and parsimony (according to Bayesian model selection criteria) as well as the best description of the data (according to PPMC). Based on the proposed model’s estimated overall class proportions, about one-half of individual item responses were “S-only class” responses, whereas about one-quarter each were “S&F class” and “F-only class” responses. Another interesting finding refers to the model’s RT results. In particular, median RTs of “F-only class” responses were about 1.1 seconds faster than “S-only class” responses, which were another 0.6 seconds faster than “S&F class” responses. These findings could reconcile some of the inconsistent results in the literature concerning the effect of faking on RTs. Specifically, the results suggest that faking does not unconditionally increase or decrease RTs. It rather seems to depend on the particular response process underlying a response that involves faking. When a test-taker retrieves an honest response and then edits it according to desirability, this seems to go along with longer RTs. In contrast, when a test-taker bypasses the retrieval of an honest response and just gives a desirable answer, this seems to be associated with shorter RTs.

To sum up, the proposed model provides a psychometrically elegant and flexible approach to modeling faking-related response strategies while making use of RTs as a form of additional behavioral data. It allows accounting for these strategies on a person-by-item level and provides a basis for a sophisticated investigation of the response process associated with faking. One important use case of the model can be the examination of what makes items especially susceptible to mere faking. The current parameterization of the model allows comparing item-specific class proportions in a fixed set of items, which can be useful for item selection purposes in a fixed item set. However, to be able to draw general conclusions about characteristics of items that make them prone to mere faking, an alternative parameterization in future work could be to replace the unconstrained item-class intercepts in the latent response model by linear combinations of item-level predictors, such as content features like relevance for the job or technical features like item length (cf. Ulitzsch, Yildirim-Erbasli, et al., 2022). If such item characteristics indeed predicted “F-only class” (or also “S&F class”) membership, this would be valuable information for test and item construction in general and could be used to develop instruments that are less susceptible to faking from the outset.

Nevertheless, some limitations of the proposed model must not be overlooked. For instance, as shown in the simulation, response strategy classifications of individual item responses are associated with considerable uncertainty, such that it is more advisable to look for general trends regarding class membership than to confidently interpret classifications of single responses. Also, the recovery of some parameters seems to deteriorate when strategy use is truly constant within test-takers. Along with that, decent class separability with respect to item responses and RTs can be expected to be important (see Pokropek, 2016; Ulitzsch et al., 2024). In addition, some of the parameter constraints in the proposed model can be questioned (e.g., ordinal latent classes, highest RTs in the “S&F class”). Future studies could drop these constraints, but convergence issues might then occur due to the increased estimation complexity. Another direction for future work would also be to further explore the utility of additional behavioral data for the modeling of faking. Process data such as eye and mouse movements or switches between response categories could provide valuable information – both from a modeling and substantive research perspective.

5 Additional Analysis: Comparison of Different Faking Models

The three articles of this dissertation present different psychometric models of faking. To compare the presented models with each other as well as with other existing faking models, I have conducted an additional analysis that goes beyond the three articles of this dissertation. In particular, the additional analysis should provide an integrative comparison of whether and to what extent the different faking models can actually improve the measurement of substantive traits. Article III might have compared person parameter recovery between the models of this dissertation in a simulation, but a comparative examination in empirical data seems warranted to have a more complete picture of the measurement properties of the different faking models. A well-suited dataset for such an analysis is the collected experimental data from Article I. This dataset contains responses from the same test-takers under an LS and HS condition including item-level RTs, allowing a comparison of how well LS substantive trait score estimates can be recovered when substantive trait scores in the HS condition are estimated using different models. Under the assumption that substantive trait score estimates in the LS condition are not systematically biased by faking, differences in LS-HS correlations between models indicate differences in how well the models' HS substantive trait score estimates represent the actual substantive trait levels (cf. Article I).

Along with the three models from the three articles of this dissertation (denoted as Model I, II, and III from now on), I considered further models in the analysis, including a full-information IRT bifactor model (Gibbons et al., 2007) as well as Brown and Böckenholt's (2022) F-GoM model with items as indicators. I additionally included a non-model-based faking correction method, in which the total number of extreme item responses in the direction of the respective trait (so-called blatant extreme responding; Landers et al., 2011; Levashina et al., 2014) is partialed from the sum score of each substantive trait scale (denoted as "residual method"). As a reference model not accounting for faking, I considered a multidimensional GPCM (MGPCM). For the analysis, I fitted all models to the data from the HS condition. To get LS substantive trait score estimates, I fitted an MGPCM to the LS condition data. Like in the other additional analysis in Section 3.1, I used the items selected by Geyer (2023; 60 items) instead of the full 125 items to avoid computational overload. I did not model response styles in any of the dissertation models to not bias comparisons to the other methods. For model fitting, I used the models' standard estimation procedures.

Table 2 shows the obtained LS-HS correlations. Several points are worth noting: First, LS-HS correlations were for all Big Five traits descriptively higher when the HS substantive trait scores were estimated using Model I than when they were estimated using a MGPCM (i.e., the reference model). Four out of five correlation differences were significant at $\alpha = .05$. Hence, as found in Article I with different item compositions and ERS being accounted for, the results indicate that the non-mixture version of the MNRM including a faking dimension (Model I) improves the measurement of substantive traits in high-stakes data compared to a regular multidimensional IRT model. The differences in correlations might not be large, but they consistently go in the same direction and, according to the simulation in Article I, larger effects are not to be expected given a low faking impact in experimental faking data as well as considerable variety in item desirability trajectories. Indeed, the item set composed by Geyer (2023) featured 51.8% items with increasing desirability trajectories, 30.4% items with an inverted-U-shaped desirability trajectory, and 17.9% items with decreasing desirability trajectories.

Table 2

Correlations of Substantive Trait Score Estimates Between the LS and HS Condition From Article I for Different Models

	MGPCM	Model I	Model II	Model III	Bifactor model	F-GoM model	Residual method
E	.524	.531*	.549† (.563**)	.539*	.241	.407	.404
A	.685	.706***	.683 (.697)	.657	.665	.373	.532
C	.584	.618***	.625* (.639***)	.591	.581	.354	.430
ES	.522	.552***	.561* (.573**)	.543*	.381	.357	.444
O	.606	.613	.653** (.667***)	.606	.541	.479	.553
averaged	.588	.608	.617 (.631)	.589	.496	.395	.475

Note. N = 1070. Low-stakes (LS) substantive trait scores were estimated using a multidimensional generalized partial credit model (MGPCM), and then correlated with the different models' high-stakes (HS) substantive trait score estimates. Correlations were averaged across the Big Five traits based on Fisher's z-transformation. Models I, II, and III, respectively, are the models from Articles I, II, and III. Values in brackets (Model II) reflect the LS-HS correlation when not considering test-takers classified into the "F-only class". F-GoM = "faking-as-grade-of-membership"; E = Extraversion; A = Agreeableness; C = Conscientiousness; ES = Emotional Stability; O = Openness.

Asterisks indicate whether an LS-HS correlation for a model and Big Five trait was significantly higher than the corresponding LS-HS correlation for the MGPCM (based on a z-test for overlapping correlations from dependent groups; Steiger, 1980): †$p < .10$, *$p < .05$, **$p < .01$, ***$p < .001$.

Second, four LS-HS correlations were also significantly higher for Model II than for the reference model when considering only the test-takers who were not classified into the “F-only class”. An exclusion of such test-takers from the analysis is sensible since their substantive trait score estimates in Model II are meaningless by the definition of the “F-only class”. When doing so, the size of LS-HS correlations increased compared to Model I, which suggests that Model II with its disentanglement of different response strategies improves measurement further.

Third, regarding Model III, only two LS-HS correlations were significantly higher than the reference model’s corresponding correlations, and the average correlation across the Big Five traits was virtually the same as for the reference model. The results thus do not provide empirical evidence that response strategy classification on the item level really improves the measurement of substantive traits. However, Bayesian model selection criteria did not prefer Model III over Model II in this dataset, suggesting that strategy use can be explained by a mere person variable in the present data. In addition, there are other possible explanations for reduced LS-HS correlations when class membership is allowed to vary within persons. Along with potentially invalid class assignments, reduced LS-HS correlations can, for instance, be an artifact of reduced reliability of substantive trait scores in such models, as responses classified into the “F-only class” reduce the number of items available for the estimation of substantive trait scores.[7] Relatedly, discarding responses from specific items implies that specific content features of the substantive traits measured by these items go missing for individual test-takers, which can attenuate correlations with LS substantive trait score estimates that are based on the full item set. Also, if test-takers are classified into the “F-only class” across all items, the Bayesian estimation of the model just samples substantive trait scores from the prior distribution for these test-takers, attenuating LS-HS correlations even further.

For the other model-based and non-model-based adjustment methods, LS-HS correlations were substantially lower than for the dissertation models and even lower than for the reference model. That is, the results strongly indicate that these methods remove substantive variance from substantive trait scores and thus do more harm than good. For the bifactor model and residual method, this is not surprising considering the mere logic of general and specific factors in bifactor models (see Section 1.2.1) as well as the fact that the residual method is very similar to partialing SDR scale scores from sum scores, which has been shown to be an

[7] Indeed, empirical reliabilities of substantive trait scores ($rel_{emp} = 1 - \frac{\overline{SE(\hat{\theta})^2}}{\sigma^2_{\hat{\theta}}}$) were lower in Model III compared to the other models. When applying a double correction for attenuation, the average LS-HS correlations were .702 for the MGPCM, .731 for Model I, .789 for Model II, and .741 for Model III. These values, however, still do not indicate a superiority of Model III regarding the measurement of substantive traits.

inappropriate faking-adjustment method (e.g., de Vries et al., 2014; Reeder & Ryan, 2011). For the F-GoM model, LS-HS correlations were especially low. However, similar caveats apply like for Model III, especially a large proportion of responses classified as faked (71.3% in the present dataset), reducing reliability and altering content features of the assessed substantive traits. Of course, the F-GoM model could also just not be directly applicable to polytomous items as indicator variables.

6 General Discussion

In this article-based dissertation, I have developed and examined psychometric latent variable models to disentangle substantive traits and faking in high-stakes personality assessments. Thereby, I have used different methods of statistical modeling, including multidimensional IRT as well as mixture models, but also met the problem of faking with item and test construction techniques. In the following, I will integrate the lessons learned by contrasting the approaches from the three articles, emphasize general limitations, discuss implications for research and practice, and derive directions for future studies.

6.1 Overview of the Models

All three models of this dissertation build on the MNRM as described by Falk and Cai (2016). Model I constitutes an application of the MNRM to the response bias of faking, with scoring weights of faking representing item- and category-specific desirability values. Model II is a confirmatory mixture extension of the MNRM to disentangle a response strategy where responses are influenced by substantive traits and faking from response strategies where responses are only influenced by substantive traits or faking. Model III further extends the model to allow for switches between strategies over the course of the assessment while making use of RTs as additional behavioral data. Common to the three models is that they yield faking-adjusted substantive trait scores and, at the same time, a measurement of faking for each test-taker. They are thus not limited to either the detection or prevention of faking, which constitutes a major advantage over many other approaches to faking (see Section 1.1.3). In addition, the three dissertation models account for item-specific, potentially nonmonotonic relationships between response categories and social desirability, which is something previous latent variable models of faking cannot afford (see Section 1.2.1). Also, they all give an indication about how faking is related to individual items, either through item slopes of faking (Models I, II, and III) or through item-class intercepts in the latent response model component of Model III. This makes them useful not only for the mere modeling of faking but also for item selection and test construction purposes. A juxtaposition of the three models, contrasting their features, ways of estimation, as well as advantages and disadvantages, is provided in Table 3.

In general, the simulation studies in this dissertation have demonstrated that all three models can recover parameters well when the data-generating process aligns with the models' assumptions. At the same time, underparameterized models (i.e., models lacking certain components of the data-generating faking process) yield substantially worse estimates for

Table 3

Overview of the Three Models of This Dissertation

	Features	Estimation	Advantages	Disadvantages
Model I (Seitz et al., 2024)	● Faking as a continuous latent variable (= degree of response alignment with social desirability) ● Assumption of a common response strategy among all test-takers ● Only item responses modeled	● MH-RM algorithm (Cai, 2010) in the R package *mirt* ● Straightforward specification of the model ● Computation time in a dataset with N = 500, I = 30, 3 substantive trait dimensions: 3.3 minutes	● Straightforward comparison of person parameters ● Response styles can easily be modeled alongside faking. ● Fast estimation, only knowledge in multidimensional IRT required	● Not accounting for response strategy heterogeneity ● Biased parameter estimates when test-takers use different response strategies
Model II (Seitz, Alagöz, & Meiser, 2025)	● Faking as both a continuous latent variable (= degree of response alignment with social desirability) and a discrete latent class (= response strategy use) ● Assumption of different response strategies across test-takers ● Only item responses modeled	● Bayesian MCMC algorithm in JAGS, via the R package *runjags* ● Own model syntax required, including the definition of priors ● Computation time in a dataset with N = 500, I = 30, 3 substantive trait dimensions: 6.3 hours	● Faking detection on the person level ● Covariates of response strategy use can be examined.	● Potentially limited comparability of person parameters across classes due to non-invariance of item-category intercepts ● Limited to the case of constant response strategy use per test-taker ● Slow estimation, susceptible to local solutions, knowledge in Bayesian estimation required
Model III (Seitz & Ulitzsch, 2025)	● Faking as both a continuous latent variable (= degree of response alignment with social desirability) and a discrete latent class (= response strategy use/inclination) ● Assumption of different response strategies across test-takers and items ● Item responses and RTs modeled	● Bayesian MCMC algorithm in JAGS, via the R package *runjags* ● Own model syntax required, including the definition of priors ● Computation time in a dataset with N = 500, I = 30, 3 substantive trait dimensions: 11.9 hours	● High flexibility → The other models can be represented as special cases. ● Modeling of RTs allows for testing substantive hypotheses on the response process. ● Items prone to particular response strategies can be identified.	● Risk of overfitting when the model is overparameterized ● Potentially limited comparability of person parameters across classes due to non-invariance of item-category intercepts ● Slow estimation, susceptible to local solutions, knowledge in Bayesian estimation required

Note. Computation times were assessed on an AMD Ryzen 7 7700 processor with 8 cores and 32 gigabytes RAM, using the simulated dataset from the first replication of the simulation study in Article III. Model I converged after 536 MH-RM iterations, Models II and III had all $\widehat{R}$ values below 1.1 after running 4000 burnin and 10000 regular MCMC iterations with 4 parallel chains. Model II was fitted without predictors of class membership. RT = response time; MH-RM = Metropolis-Hastings Robbins-Monro; MCMC = Markov chain Monte Carlo.

almost all parameters. Vice versa, in case of overparameterization (e.g., no faking in the data, sample composed of only one class), Models I and II can still estimate parameters without bias and without reduced accuracy. Model III, in contrast, seems to yield slightly less accurate estimates when response strategy use truly does not vary within persons, even though the model is able to approximate constant class membership per person. Furthermore, the additional analysis in Section 5 has provided evidence that Models I and II also empirically improve the measurement of substantive traits. With regard to Model III, the additional analysis has not provided support for this. However, there has been no indication that either of the dissertation models systematically overextracts substantive variance in the sense that the estimation of substantive trait scores becomes worse than in models not accounting for faking, which is the case in existing faking models that make oversimplified assumptions. For the sake of model parsimony, it is nevertheless advisable to always check whether the additional complexities of the models are empirically justified. As demonstrated in the articles, model selection criteria, estimated class proportions, and model-implied class memberships across items can be considered to decide which model to choose.

6.2 General Limitations

Since all three models of this dissertation are based on the MNRM, there are a couple of limitations and drawbacks that Models I to III have in common. Most of them are mentioned in the three articles, but the most important ones deserve a closer look here. To begin with, an interpretational caveat that applies to all three models must be noted. Namely, a test-taker's faking score in the MNRM is only a relative description of his or her faking behavior compared to the other modeled test-takers, such that person parameters of faking must be interpreted normatively (see Bolt & Meng, 2025, for the same interpretational caveat when modeling response styles). This is due to the mere fact that faking is modeled as a latent variable, which does not have a natural origin. The scale indeterminacy of latent variables is usually resolved by fixing the latent mean to an arbitrary number, typically 0. Hence, a faking score of 0 only indicates an average faking level and not the absence of faking in absolute terms. Correspondingly, negative faking scores indicate below-average faking levels and not socially *un*desirable responding in terms of faking bad. The absence of faking in the response process is rather captured by item slopes of faking being 0, implying that the faking dimension does not explain variance in item responses (cf. Articles II and III).

Another aspect not to be overlooked is the fact that the MNRM approach to modeling faking requires a-priori information on the desirability of all item-category combinations before the model can be applied to data. In practice, this means that pilot data needs to be collected. Running a pilot study requires resources if it is conducted thoroughly in order to obtain desirability values that are representative of the social context of the actual assessment. At the same time, collected pilot data is usually not transferable to other empirical datasets with different items or from different assessment contexts. However, considering that faking is a complex interplay of person and situation characteristics (e.g., Kuncel & Tellegen, 2009; Tett & Simonet, 2011), the specificity of desirability values just seems to be necessary to capture faking adequately. Also, in an applied selection setting, the costs of hiring the wrong candidate because of poor personality assessment will probably outweigh the costs of running a pilot study tailored to the selection context at hand.

Whereas the two previous aspects can be considered as an interpretational and pragmatic downside of the MNRM, respectively, a more important point is the person-invariant specification of scoring weights of faking. With such a specification, it is implicitly assumed that all test-takers perceive social desirability equivalently. However, as can be seen in the error bars in Figure 1, people differ in their ratings of desirability. Hence, the question can be raised whether this way of modeling faking is susceptible to interindividual differences in desirability perceptions. The articles of this dissertation do not provide an answer to this question, but a recent simulation study on this issue has provided first evidence for considerable robustness of the MNRM against violations of person-invariant scoring weights of faking (Kleinbub & Seitz, 2025).[8] In this study, the recovery of substantive trait scores was not impaired when the simulated test-takers' individual desirability perceptions fluctuated unsystematically around the average desirability trajectories (see Figure 4a). Only when there were large subgroups of simulated test-takers whose desirability perceptions deviated systematically from the assumed desirability trajectories, person parameter recovery deteriorated slightly, but the model was still clearly superior to a model in which faking was not accounted for (see Figure 4b).

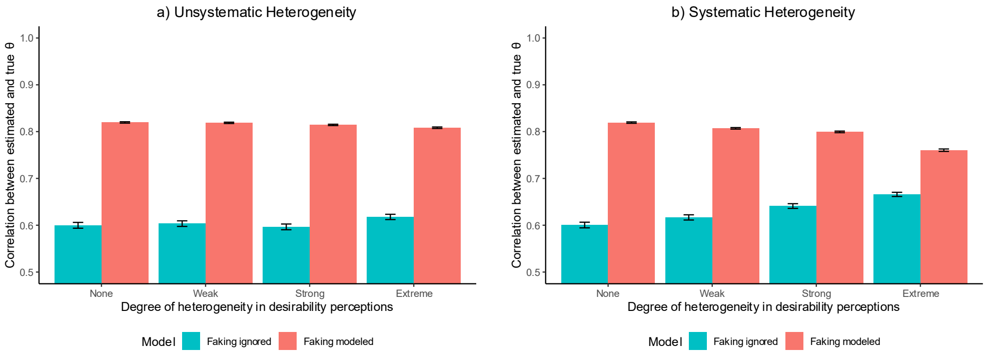
Furthermore, there is a limitation that concerns Models II and III, namely the estimation complexity of these models using Bayesian MCMC sampling. Let alone that many applied researchers or practitioners inclined to use one of the models probably do not know much about Bayesian estimation techniques and associated software packages, the two models require a lot of computational resources, both in terms of computation time and computation power. For example, in a dataset of $N = 500$ persons and $I = 30$ items, Models II and III, respectively, take around 6 and 12 hours to estimate on a powerful desktop computer while requiring 2.4 and 4.6 gigabytes of RAM (see the note of Table 3 for more information). With increasing

[8] This work was based on the first author's master thesis under my supervision.

computational power, this practical limitation might not be a big issue in the future. But even if all computational resources are available, Models II and III are still hard to estimate because they constitute complex mixture models. Mixture models in general are difficult to estimate because of multimodal likelihood functions (e.g., Hipp & Bauer, 2006; McLachlan & Peel, 2000), and the multidimensional nature as well as nominal treatment of response categories obviously do not make the estimation easier. A consequence of multimodal likelihood functions is that the estimation is prone to finding local solutions. That is, when starting the estimation with random initial values, the different MCMC chains might yield systematically diverging parameter estimates, especially in empirical data. To alleviate this problem, one can employ an estimation procedure that optimizes initial values (cf. O'Hagan et al., 2012). In particular, a large number of parallel chains with random initial values can be run before taking the parameter estimates from the chain yielding the highest model likelihood as initial values (with additional random noise) for the actual model estimation. With that, the models yielded satisfactory convergence in Articles II and III. There is nevertheless no guarantee that this procedure really finds the global maximum of the likelihood function – but most numerical approximation methods do not guarantee this.

Figure 4

Recovery of Substantive Trait Scores in the Simulation Study From Kleinbub & Seitz (2025)



Note. The depicted recovery of substantive trait scores applies to the case of 6 items per substantive trait scale, a sample size of 500, and a strong faking impact in the data. In the data generation, the degree of heterogeneity in desirability perceptions was operationalized through the dispersion of individual faking scoring weights (unsystematic heterogeneity) and through the group size of simulated test-takers with systematically deviating faking scoring weights (systematic heterogeneity). Models ignoring faking only included dimensions for substantive traits, whereas models accounting for faking also included a faking dimension. Values reflect the back-transformed mean of the Fisher-*z*-transformed correlations between estimated and true person parameters across replications within a condition. Error bars represent the standard error of the mean.

6.3 Implications for Research and Practice

What are the implications of this dissertation for research and practice? Since faking entails a deliberate response distortion, accounting for faking is apparently more relevant when assessment results have important consequences for test-takers compared to contexts where test-takers are anonymous and stakes are low. Nevertheless, the models of this dissertation are not only valuable for applied high-stakes assessments, but also for research. As opposed to faking prevention methods, the presented models do allow for substantive investigations of faking and the associated response process. For instance, all three models allow researchers to study relationships of a continuous faking variable with the modeled substantive personality traits or with external variables like intelligence or performance measures. Likewise, Models II and III also allow studying covariates of faking-related response strategy use. Thus, the models offer researchers who wish to better understand the substantive nature of the faking construct an appealing framework for more sophisticated examinations that would not be possible with traditional methods like SDR scales. Model III additionally provides an approach for psychometric research to systematically examine what makes items prone to certain response strategies, which can help to make rating scale measures less fakable from the outset. The proposed item refinements in Article I can also be a leverage point for less fakable rating scale measures. However, such item refinements are not to be recommended without reservation because subtle changes in the meaning of the assessed construct can occur when deconfounding descriptive and evaluative aspects of item content. As noted in Article I, if the goal is to have a fair assessment that is not contaminated by faking, the proposed item refinements will be more appropriate than in contexts like personality research, where the goal is to measure narrowly defined traits.

With regard to applied assessment contexts such as personnel selection, the features of the models – especially the adjustments of substantive trait scores as well as the measurement of faking – speak for themselves. However, from my experience working as a freelancer in the testing industry for a couple of years, I know that psychometric quality alone does not imply value in applied diagnostic settings. There are several obstacles that need to be overcome for a psychometric model to be applied in practice. Along with estimation complexity (see Section 6.2), a major challenge for practitioners is the scoring of new cases using a latent variable model. Scoring new cases means transforming raw scores into standardized scores with respect to a standardization sample. In practice, raw scores are usually sum scores, because they are easy to compute and do not depend on item parameters. Computing test scores using a latent variable model, in contrast, is considerably more complex and requires a sufficiently large sample to reliably estimate model parameters. However, given that a large standardization sample is available, one can calibrate (i.e., estimate) item parameters based on the standardization sample and then treat them as fixed, such that only person parameters need to be estimated when new cases are successively coming in. This estimation then only takes a few seconds, even in a Bayesian context, and the R package *mirt* (Chalmers, 2012) even offers a function for that. Hence, I would argue that it is without a doubt possible to implement a latent variable model of faking in applied assessment contexts. From a mere psychometric point of view, this is advisable because test-takers with the same true substantive trait level but different faking levels would otherwise have different expected test scores, which is an issue of both construct validity and test fairness. Based on the findings of the three articles and the additional analysis in Section 5, at least Model I can be recommended to be used – also considering its vast speed advantage in model estimation. Model II also performed well, but the potentially limited comparability of person parameters across classes due to the non-invariance of item-category intercepts must be considered. Model III requires more research to make a clear statement about whether it can really be recommended for use in applied diagnostics.

Nevertheless, there are also other aspects that will determine if latent variable models of faking in general will eventually be used in practice. For example, it will strongly depend on the extent to which the respective method is accepted by different stakeholders. Some might approve of it because of its psychometric features outlined above. Others might react skeptically, be it because the computation of test scores is not as intuitive as in the case of sum scores or because of legal concerns associated with a model-based adjustment of scores. To circumvent the issue of post-hoc adjustments of scores, approaches that aim at preventing faking in the first place, such as MFC tests, can be considered as a viable alternative to self-report tests using rating scales. The elimination of faking undoubtedly has high appeal for high-stakes assessments, however, the pitfalls mentioned in Section 1.1.3 must be overcome for MFC measures to offer a real advantage over rating scale measures. This especially concerns the derivation of normative as opposed to ipsative test scores. With the application of TIRT, it is technically possible to obtain normative scores (Brown & Maydeu-Olivares, 2011, 2013). However, just like for the models of this dissertation, a scoring technique based on a complex latent variable model is required. Also, appropriate item blocks need to be composed. This constitutes a big challenge because items within blocks, on the one hand, should be desirability-matched to reduce fakability (Cao & Drasgow, 2019; Wetzel et al., 2021) and, on the other hand, should not all be keyed in the same direction to facilitate or even allow the estimation of normative scores (Bürkner et al., 2019; Frick et al., 2023). Due to the confound of descriptive and evaluative aspects of item content (cf. Article I; Peabody, 1967), these two requirements usually contradict each other in the sense that most positively-keyed items are generally desirable (undesirable) and most negatively-keyed items are generally undesirable (desirable). A recent study has shown that it can be possible to strike a balance between desirability-matched and mixed-keyed item blocks to reduce fakability and, at the same time, yield adequate properties of test scores (Li et al., 2025; see also Holling, 2025). Nevertheless, optimal MFC test construction remains a challenge (see Schünemann, 2025).

Time will tell whether statistical models of faking in rating scale data or MFC measures will find their way to applied diagnostics and consistently replace methods from classical test theory. I would expect that, with the increasing popularity of machine learning and artificial intelligence (AI) in everyday life, sophisticated statistical models in personality assessment will also gain more acceptance among practitioners. Above all, however, it will depend on which method is of highest utility for practitioners. As elaborated above, faking distorts construct validity of self-report measures, and the models presented in this dissertation seem to be able to improve construct validity compared to models not accounting for faking. Nonetheless, even if measures then better capture the psychological constructs intended to be assessed, I know from my experience in the testing industry that this psychometric feature will interest practitioners less than the effects on criterion-related validity, especially when the assessment is part of vocational aptitude testing in hiring contexts.

Conceptually, adjusting substantive trait scores for faking can affect criterion-related validity in three different ways. First, the model's adjustments can improve the prediction of criteria. Such an effect would suggest that faking acts as a suppressor variable (Bing et al., 2011; Hakstian & Ng, 2005), and would align with Speer et al.'s (2023) recent meta-analytic finding of higher predictive validity in high-faking contexts when personality measures are faking-resistant. Second, the model's adjustments can have a null effect. This would be in line with research suggesting that faking does not harm prediction (e.g., Morgeson et al., 2007; Ones et al., 1996, 2007; Paunonen & LeBel, 2012). Third, the model's adjustments can reduce predictive validity. This effect would fit with the claim that faking signals social skills (e.g., Marcus, 2009) and reflects a variable that in itself has predictive power, such that correlations between self-reported personality and outcomes decline when faking is controlled for (cf. Li & Bagger, 2006). Even though the prediction of outcomes is a crucial aspect in diagnostics, I would argue that the quality of a psychometric model should primarily be judged based on considerations of construct validity as opposed to criterion-related validity (cf. Section 1.1.2). This assertion mainly rests on the fact that, when modeling response biases like faking, the effects of the modeling on criterion-related validity strongly depend on the true relationships of the response bias with substantive traits and the criterion variable at hand (Komar et al., 2008). These relationships are highly context-specific and, per se, independent of the mere psychometric quality of the model. For example, in jobs where impression management is objectively advantageous for better performance, faking might correlate positively with performance outcomes (e.g., number of products sold in a sales job), such that faking adjustments can be expected to reduce predictive validity. Conversely, in jobs where high integrity is of importance, faking might correlate negatively with the relevant outcome variables (e.g., responsible behavior in a law enforcement context), such that increases in predictive validity through faking adjustments can be expected.

6.4 Directions for Future Research

Even though there are already a lot of psychometric models of faking, the research projects of this dissertation warrant further work on the presented models, including validation studies and applications to other contexts than high-stakes assessments. Regarding the models themselves, there are some appealing further developments that future research could examine. First, as mentioned above, faking scores in the MNRM are relative quantifications of the faking degree due to the scale indeterminacy of latent variables. However, it would be appealing to be able to assess the faking level of a test-taker in an absolute manner. Also, the assumption of a normally distributed faking variable might not describe actual faking behavior best. It would hence be interesting to model a different distributional form of faking where a faking score of 0 also implies the absence of faking. This would correspond to modeling faking as a unipolar as opposed to a bipolar construct. Faking could, for instance, be modeled using a log-normal distribution, which is only defined for positive values (see Huang & Bolt, 2024; Lucke, 2015, for existing unipolar IRT approaches). Second, one could test whether a response strategy where item responses are only driven by social desirability (cf. the “F-only class” in Articles II and III) is better described by a model of stochastic independence instead of a one-dimensional faking model. In such a model of independence, faking would only be captured by item-category intercepts that represent the desirability of a response category at a particular item, while all variation in this class would be unsystematic. Third, according to Falk and Cai (2016), the MNRM technically does not require fixed scoring weights of response bias dimensions but allows for estimating these values when appropriate constraints are imposed. It would hence be appealing to freely estimate scoring weights of faking instead of fixing them to assumed desirability values (cf. Bolt & Johnson, 2009, for an estimation of scoring weights of response styles). This would alleviate the burden of having to run a pilot study in advance. One could also consider estimating scoring weights in a random-effects or even mixture framework, which would offer a different opportunity to explore parameter heterogeneity. However, one would first have to investigate thoroughly whether the model still captures faking when the scoring weights are estimated instead of fixed.

Despite these interesting model development ideas, I would advise against further model extensions at the next step; firstly because the models already are very complex, secondly because the field would probably benefit more from validation evidence than from further faking models. An appealing validation study would be to investigate how well the models of this dissertation as well as other faking models can mitigate the effects of faking compared to MFC measures, which are becoming increasingly popular. The low fakability of MFC measures has primarily been found in comparison to rating scale measures without faking being statistically accounted for. Results may well look different when psychometric models of faking are considered as comparison models. In such a validation study, fakability should also not only be operationalized through the extent of mean shifts between an LS and HS condition. An arguably more valid fakability operationalization would instead be how well estimates of substantive trait scores in an HS condition represent actual substantive trait levels (cf. Schulte et al., 2024). LS-HS correlations would be one option to examine this (see Article I and the additional analysis in Section 5). As LS substantive trait score estimates might also not be free of SDR, another option would be to use other measures of personality as a comparison standard, such as measures that do not entail self-report (e.g., observer ratings of personality). Even though these personality measures also come with limitations (Connolly et al., 2007; König et al., 2017), such validation evidence would be valuable to guide follow-up research.

Along with validation studies, further directions for future research are applications of the models of this dissertation to other fields in the social sciences. The models could, for instance, be applied in the context of survey methodology. Among other things, one could investigate SDR and faking depending on different data collection modes. Item slopes of faking in the MNRM could thereby be used to compare the faking susceptibility of telephone versus face-to-face interviews or personalized versus anonymous surveys. When enough data is available, Model III could also be used to identify questions that are prone to dishonest responding irrespective of the data collection mode, which could improve the design of existing and new surveys. Additionally, it would be interesting to apply the models to forensic assessment contexts, where the faking dimension would represent *faking bad* as opposed to *faking good*. Applications to different low-stakes datasets would also be valuable, as one could address the question of whether the MNRM allows a proper modeling of self-deception, too, or rather captures substantive variance in low-faking contexts. Moreover, it would be important to investigate cross-cultural differences in faking. Such differences would be interesting from a substantive perspective, but also from a measurement perspective. Namely, if the response process associated with faking differed qualitatively between cultures and, as a consequence, measurement invariance with respect to the faking dimension in the MNRM did not hold, this would call for caution when comparing test-takers with different cultural backgrounds.

Finally, future research must not ignore advancements in machine learning and AI, which have led to increasingly powerful prediction models in recent years. Combining traditional psychometric methods with recent machine-learning-based methods could hence yield more accurate predictions of outcome variables while still ensuring a sound measurement of psychological constructs. At the same time, mere faking detection (see Lee & Ziegler, 2025; Nie et al., 2025; Röhner et al., 2022) might improve steadily in the future, which could open up new possibilities for dealing with faking in research and practice. Machine learning and AI methods could also be leveraged to develop new personality assessment tools that are less susceptible to faking from the outset. There are already several data-driven attempts to measure personality beyond the conventional self-report questionnaires, including the use of chatbots (Fan et al., 2023), video interviews (Hickman et al., 2022; Koutsoumpis et al., 2024), virtual reality environments (Vargas et al., 2024), as well as social media data and other digital footprints (Fernandez et al., 2021; Mönke et al., 2024). However, the vulnerability of these approaches to deliberate response distortions is still unexplored. At the same time, the question remains if these approaches really measure the personality constructs they claim to measure.

6.5 Conclusion

In this dissertation, I have investigated the response bias of faking in high-stakes personality assessments by means of psychometric modeling. Across three articles and additional analyses in this synopsis, I have demonstrated the usefulness of different multidimensional IRT models when dealing with desirability-related response distortions, and have shown how appropriate item construction can make an important contribution. The models of this dissertation treat faking as a continuous latent variable using the MNRM and allow for modeling qualitatively different faking-related response strategies using mixture modeling techniques. By making use of item RTs, the most complex model proposed in this dissertation can also account for switches between response strategies over the course of the assessment. Beyond these methodological contributions, the presented models offer a framework for researchers to better understand the substantive nature of the faking construct. This dissertation underlines that faking is not an intractable response bias but that the appropriate application of psychometric models can help to better manage its adverse effects. However, further research is required to examine how the presented models fare against alternative approaches to dealing with faking (e.g., MFC tests) and what practical benefits the models can bring in applied diagnostic settings.

Bibliography

Alagöz, Ö. E. C., & Meiser, T. (2024). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement, 84*(5), 957–993. https://doi.org/10.1177/00131644231206765

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing* (4th ed.). American Educational Research Association.

Atherton, O. E., Robins, R. W., Rentfrow, P. J., & Lamb, M. E. (2014). Personality correlates of risky health outcomes: Findings from a large Internet study. *Journal of Research in Personality, 50*, 56–60. http://doi.org/10.1016/j.jrp.2014.03.002

Austin, E. J., Deary, I. J., & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality and Individual Differences, 40*(6), 1235–1245. https://doi.org/10.1016/j.paid.2005.10.018

Bäckström, M., & Björklund, F. (2024). Why forced-choice and Likert items provide the same information on personality, including social desirability. *Educational and Psychological Measurement, 84*(3), 549–576. https://doi.org/10.1177/00131644231178721

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*(3), 335–344. https://doi.org/10.1016/j.jrp.2008.12.013

Bäckström, M., Björklund, F., Maddux, R. E., & Lindén, M. (2023). The NB5I: A full-scale Big-Five inventory with evaluatively neutralized items. *European Journal of Psychological Assessment, 39*(2), 132–140. https://doi.org/10.1027/1015-5759/a000687

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1–26. https://doi.org/10.1111/j.1744-6570.1991.tb00688.x

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

Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment, 31*(4), 532–544. https://doi.org/10.1037/pas0000619

Bing, M. N., Kluemper, D., Kristl Davison, H., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes, 116*(1), 148–162. https://doi.org/10.1016/j.obhdp.2011.05.006

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*(3), 478–494. https://doi.org/10.1037/0021-9010.74.3.478

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. https://doi.org/10.1007/bf02291411

Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. *Psychometrika, 79*(3), 515–537. https://doi.org/10.1007/s11336-013-9390-9

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

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*(5), 335–352. https://doi.org/10.1177/0146621608329891

Bolt, D. M., & Meng, L. (2025). IRT-based response style models and related methodology: Review and commentary. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. https://doi.org/10.1111/bmsp.70006

Borkenau, P., Zaltauskas, K., & Leising, D. (2009). More may be better but there may be too much: Optimal trait level and self-enhancement bias. *Journal of Personality, 77*(3), 825–858. https://doi.org/10.1111/j.1467-6494.2009.00566.x

Brown, A. (2010). *How item response theory can solve problems of ipsative data* [Doctoral dissertation, University of Barcelona]. https://www.tesisenred.net/handle/10803/80006

Brown, A, & Bartram, D. (2009, April 2–4). *Doing less but getting more: Improving forced-choice measures with IRT* [Paper presentation]. 24th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, United States.

Brown, A., & Böckenholt, U. (2022). Intermittent faking of personality profiles in high-stakes assessments: A grade of membership analysis. *Psychological Methods, 27*(5), 895–916. https://doi.org/10.1037/met0000295

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing rater biases in 360-degree feedback by forcing choice. *Organizational Research Methods, 20*(1), 121–148. https://doi.org/10.1177/1094428116668036

Brown, A., & Maydeu-Olivares, A. (2011). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement, 71*(3), 460–502. https://doi.org/10.1177/0013164410375112

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36–52. https://doi.org/10.1037/a0030641

Bürkner, P.-C., Schulte, N., & Holling, H. (2019). On the statistical and practical limitations of Thurstonian IRT models. *Educational and Psychological Measurement, 79*(5), 827–854. https://doi.org/10.1177/0013164419832063

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

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

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80*(1), 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*(1), 88–106. https://doi.org/10.1037/1040-3590.19.1.88

Christiansen, N. D., Robie, C., Burns, G. N., Loy, R. W., Speer, A. B., & Jacobs, R. R. (2021). Effects of applicant response distortion on the relationship between personality trait scores and cognitive ability. *Personality and Individual Differences, 171*, 110542. https://doi.org/10.1016/j.paid.2020.110542

Christiansen, N. D., Robie, C., Burns, G. N., & Speer, A. B. (2017). Using item-level covariance to detect response distortion on personality measures. *Human Performance, 30*(2–3), 116–134. https://doi.org/10.1080/08959285.2017.1319366

Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta-analytic review. *International Journal of Selection and Assessment, 15*(1), 110–117. https://doi.org/10.1111/j.1468-2389.2007.00371.x

Cook, R., Roulin, N., & Joy, K. (2024). Development, validation, and faking-resistance of an implicit measure of psychopathy in the workplace. *Human Performance, 37*(5), 245–279. https://doi.org/10.1080/08959285.2024.2422341

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C., & John, O. P. (2016). Die deutsche Version des Big Five Inventory 2 (BFI-2) [The German version of the Big Five Inventory 2 (BFI-2)]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)* [*Compilation of items and scales for the social sciences (ZIS)*]. https://doi.org/10.6102/zis247

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C. J., & John, O. P. (2019). Das Big Five Inventar 2: Validierung eines Persönlichkeitsinventars zur Erfassung von 5 Persönlichkeitsdomänen und 15 Facetten [The Big Five Inventory 2: Validation of a personality inventory for measuring 5 personality domains 15 facets]. *Diagnostica, 65*(3), 121–132. https://doi.org/10.1026/0012-1924/a000218

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression management as an expression of honesty-humility. *Assessment, 21*(3), 286–299. https://doi.org/10.1177/1073191113504619

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*(9). https://doi.org/10.18637/jss.v071.i09

Diekmann, J., & König, C. J. (2015). Personality testing in personnel selection: Love it? Leave it? Understand it! In I. Nikolaou & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment* (pp. 129–147). Psychology Press.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81–106. https://doi.org/10.1207/S15327043HUP1601_4

Dunlop, P. D., Holtrop, D., Ashby, L. M., Bharadwaj, A., & Donovan, J. J. (2022). Valence, instrumentality, expectancy, and ability as determinants of faking, and the effects of faking on criterion-related validity. *Journal of Business and Psychology, 37*(6), 1215–1233. https://doi.org/10.1007/s10869-022-09797-0

Dunlop, P. D., Telford, A. D., & Morrison, D. L. (2012). Not too little, but not too much: The perceived desirability of responses to personality items. *Journal of Research in Personality, 46*(1), 8–18. https://doi.org/10.1016/j.jrp.2011.10.004

Dunlop, P. D., Xia, M. (R.), & Anglim, J. (2025). Faking on personality assessments in high-stakes settings: A critical review. *Current Opinion in Psychology, 65*, 102057. https://doi.org/10.1016/j.copsyc.2025.102057

Ellingson, J. E., & McFarland, L. A. (2011). Understanding faking behavior through the lens of motivation: An application of VIE theory. *Human Performance, 24*(4), 322–337. https://doi.org/10.1080/08959285.2011.597477

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*(2), 155–166. https://doi.org/10.1037/0021-9010.84.2.155

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

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology, 97*(4), 866–880. https://doi.org/10.1037/a0026655

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology, 108*(8), 1277–1299. https://doi.org/10.1037/apl0001082

Feeney, J. R., Goffin, R. D., & Beshai, S. (2023). Applicant faking warnings: Are they really effective? *Personality and Individual Differences, 200*, 111899. https://doi.org/10.1016/j.paid.2022.111899

Fernandez, S., Stöcklin, M., Terrier, L., & Kim, S. (2021). Using available signals on LinkedIn for personality assessment. *Journal of Research in Personality, 93*, 104122. https://doi.org/10.1016/j.jrp.2021.104122

Frick, S. (2022). Modeling faking in the multidimensional forced-choice format: The faking mixture model. *Psychometrika, 87*(2), 773–794. https://doi.org/10.1007/s11336-021-09818-6

Frick, S., Brown, A., & Wetzel, E. (2023). Investigating the normativity of trait estimates from multidimensional forced-choice data. *Multivariate Behavioral Research, 58*(1), 1–29. https://doi.org/10.1080/00273171.2021.1938960

Fuechtenhans, M., & Brown, A. (2023). How do applicants fake? A response process model of faking on multidimensional forced-choice personality assessments. *International Journal of Selection and Assessment, 31*(1), 105–119. https://doi.org/10.1111/ijsa.12409

Furr, R. M. (2021). *Psychometrics: An introduction* (4th ed.). Sage.

Geyer, N. (2023). *Reduktion sozial erwünschten Antwortverhaltens im Big Five Inventar 2 durch Veränderung des evaluativen Inhalts der Items* [*Reduction of socially desirable response behavior in the Big Five Inventory 2 through change of the evaluative concent of the items*] [Unpublished bachelor thesis]. University of Mannheim.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*(1), 4–19. https://doi.org/10.1177/0146621606289485

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology / Psychologie canadienne, 50*(3), 151–160. https://doi.org/10.1037/a0015946

Goldammer, P., Stöckli, P. L., Escher, Y. A., Annen, H., & Jonas, K. (2024). On the utility of indirect methods for detecting faking. *Educational and Psychological Measurement, 84*(5), 841–868. https://doi.org/10.1177/00131644231209520

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*(3), 341–355. https://doi.org/10.1108/00483480710731310

Griffith, R. L., & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0018

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology, 1*(3), 308–311. https://doi.org/10.1111/j.1754-9434.2008.00053.x

Habashi, M. M., Graziano, W. G., & Hoover, A. E. (2016). Searching for the prosocial personality: A Big Five approach to linking personality and prosocial behavior. *Personality and Social Psychology Bulletin, 42*(9), 1177–1192. https://doi.org/10.1177/0146167216652859

Hakstian, A. R., & Ng, E.-L. (2005). Employment-related motivational distortion: Its nature, measurement, and reduction. *Educational and Psychological Measurement, 65*(3), 405–441. https://doi.org/10.1177/0013164404267293

Hendy, N., Krammer, G., Schermer, J. A., & Biderman, M. D. (2021). Using bifactor models to identify faking on Big Five questionnaires. *International Journal of Selection and Assessment, 29*(1), 81–99. https://doi.org/10.1111/ijsa.12316

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

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology, 107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods, 11*(1), 36–53. https://doi.org/10.1037/1082-989X.11.1.36

Hogan, R., & Blickle, G. (2013). Socioanalytic theory. In N. D. Christiansen & R. P. Tett (Eds.), *Handbook of personality at work*. Routledge.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63*(2), 272–279. https://doi.org/10.1037/0022-3514.63.2.272

Holling, H. (2025, July 22–25). *Optimal design in linear paired comparisons for Thurstonian IRT models* [Oral presentation]. 11th European Congress of Methodology, Tenerife, Spain.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*(2), 161–172. https://doi.org/10.1177/0146167203259930

Hu, J., & Connelly, B. S. (2021). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. *International Journal of Selection and Assessment, 29*(3–4), 412–426. https://doi.org/10.1111/ijsa.12338

Huang, Q., & Bolt, D. M. (2024). Unipolar IRT and the Author Recognition Test (ART). *Behavior Research Methods, 56*(6), 5406–5423. https://doi.org/10.3758/s13428-023-02275-2

Jackson, D. N., & Messick, S. (1958). Content and style in personality assessment. *Psychological Bulletin, 55*(4), 243–252. https://doi.org/10.1037/h0045996

Jebb, A. T., Ng, V., & Tay, L. (2021). A review of key Likert scale development advances: 1995–2019. *Frontiers in Psychology, 12*, 637547. https://doi.org/10.3389/fpsyg.2021.637547

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*(1), 92–114. https://doi.org/10.3102/1076998609340529

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87*(3), 530–541. https://doi.org/10.1037/0021-9010.87.3.530

Kiefer, C., & Benit, N. (2016). What is applicant faking behavior? A review on the current state of theory and modeling techniques. *Journal of European Psychology Students, 7*(1), 9–19. https://doi.org/10.5334/jeps.345

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25*(4), 273–302. https://doi.org/10.1080/08959285.2012.703733

Kleinbub, J. D., & Seitz, T. (2025, September 29–October 1). *Unmasking the faker: Heterogeneous perception of social desirability in context of the multidimensional nominal response model* [Poster]. 17th Meeting of the Methods and Evaluation Division of the German Psychological Society (DGPs), Berlin, Germany.

Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work: The role of candidates' ability to identify criteria. *Organizational Psychology Review, 1*(2), 128–146. https://doi.org/10.1177/2041386610387000

Koch, T., Holtmann, J., Bohn, J., & Eid, M. (2018). Explaining general and specific factors in longitudinal, multimethod, and bifactor models: Some caveats and recommendations. *Psychological Methods, 23*(3), 505–523. https://doi.org/10.1037/met0000146

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93*(1), 140–154. https://doi.org/10.1037/0021-9010.93.1.140

König, C. J., Steiner Thommen, L. A., Wittwer, A., & Kleinmann, M. (2017). Are observer ratings of applicants' personality also faked? Yes, but less than self-reports. International *Journal of Selection and Assessment, 25*(2), 183–192. https://doi.org/10.1111/ijsa.12171

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*(5), 768–821. https://doi.org/10.1037/a0020327

Koutsoumpis, A., Ghassemi, S., Oostrom, J. K., Holtrop, D., Van Breda, W., Zhang, T., & de Vries, R. E. (2024). Beyond traditional interviews: Psychometric analysis of asynchronous video interviews for personality and interview performance evaluation using machine learning. *Computers in Human Behavior, 154*, 108128. https://doi.org/10.1016/j.chb.2023.108128

Krammer, G., Sommer, M., & Arendasy, M. E. (2017). The psychometric costs of applicants' faking: Examining measurement invariance and retest correlations across response conditions. *Journal of Personality Assessment, 99*(5), 510–523. https://doi.org/10.1080/00223891.2017.1285781

Kuncel, N. R., & Borneman, M. J. (2007). Toward a new method of detecting deliberately faked personality tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment, 15*(2), 220–231. https://doi.org/10.1111/j.1468-2389.2007.00383.x

Kuncel, N. R., Goldberg, L. R., & Kiger, T. (2011). A plea for process in personality prevarication. *Human Performance, 24*(4), 373–378. https://doi.org/10.1080/08959285.2011.597476

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x

Kuric, E., Demcak, P., Smrecek, P., & Spilakova, B. (2025). User modeling for detecting faking-good intent in online personality questionnaires in the wild based on mouse dynamics. *Multimedia Tools and Applications, 84*(34), 43395–43431. https://doi.org/10.1007/s11042-025-20852-9

LaHuis, D. M., Blackmore, C. E., Bryant-Lees, K. B., & Delgado, K. (2019). Applying item response trees to personality data in the selection context. *Organizational Research Methods, 22*(4), 1007–1018. https://doi.org/10.1177/1094428118780310

LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods, 12*(2), 296–319. https://doi.org/10.1177/1094428107302903

Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*(1), 202–210. https://doi.org/10.1037/a0020375

Landers, R. N., & Sanchez, D. R. (2022). Game-based, gamified, and gamefully designed assessments for employee selection: Definitions, distinctions, design, and validation. *International Journal of Selection and Assessment, 30*(1), 1–13. https://doi.org/10.1111/ijsa.12376

Lee, P., Joo, S.-H., & Jia, Z. (2022). Opening the black box of the response process to personality faking: An application of item response tree models. *Journal of Business and Psychology, 37*(6), 1199–1214. https://doi.org/10.1007/s10869-022-09791-6

Lee, P., Son, M., Zhou, S., Joo, S., Jia, Z., & Cheng, V. (2025). The journey of forced choice measurement over 80 years: Past, present, and future. *Organizational Research Methods, 28*(4), 680–722. https://doi.org/10.1177/10944281251350687

Lee, T. C., & Ziegler, M. (2025). Leveraging deep learning for the detection of socially desirable tendencies in personnel selection: A proof-of-concept. *PLOS ONE, 20*(8), e0329205. https://doi.org/10.1371/journal.pone.0329205

Leite, W. L., & Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research, 45*(2), 271–293. https://doi.org/10.1080/00273171003680245

Leng, C.-H., Huang, H.-Y., & Yao, G. (2020). A social desirability item response theory model: retrieve–deceive–transfer. *Psychometrika, 85*(1), 56–74. https://doi.org/10.1007/s11336-019-09689-y

Levashina, J., Weekley, J. A., Roulin, N., & Hauck, E. (2014). Using blatant extreme responding for detecting faking in high-stakes selection: Construct validity, relationship with general mental ability, and subgroup differences. *International Journal of Selection and Assessment, 22*(4), 371–383. https://doi.org/10.1111/ijsa.12084

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*(2), 131–141. https://doi.org/10.1111/j.1468-2389.2006.00339.x

Li, M., Zhang, B., Li, L., Sun, T., & Brown, A. (2025). Mixed-keying or desirability-matching in the construction of forced-choice measures? An empirical investigation and practical recommendations. *Organizational Research Methods, 28*(2), 296–329. https://doi.org/10.1177/10944281241229784

Loy, R. W., Christiansen, N. D., Tett, R. P., Klein, K., & Toich, M. (2025). Personality test validity differs between low-stakes and high-stakes employment settings. *International Journal of Selection and Assessment, 33*(3), e70018. https://doi.org/10.1111/ijsa.70018

Lucke, J. F. (2015). Unipolar item response models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 290–302). Routledge.

Marcus, B. (2009). 'Faking' from the applicant's perspective: A theory of self-presentation in personnel selection settings. *International Journal of Selection and Assessment, 17*(4), 417–430. https://doi.org/10.1111/j.1468-2389.2009.00483.x

Marcus, B. (2022). "Faking" is neither good nor bad, it is a misleading concept: A reply to Tett and Simonet (2021). *Personnel Assessment and Decisions, 8*(1), 35–42. https://doi.org/10.25035/pad.2022.01.004

Martínez, A., & Salgado, J. F. (2021). A meta-analysis of the faking resistance of forced-choice personality inventories. *Frontiers in Psychology, 12*, 732241. https://doi.org/10.3389/fpsyg.2021.732241

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/bf02296272

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

Mazza, C., Ceccato, I., Cannito, L., Monaro, M., Ricci, E., Bartolini, E., Cardinale, A., Di Crosta, A., Cardaioli, M., La Malva, P., Colasanti, M., Tambelli, R., Giromini, L., Palumbo, R., Palumbo, R., Di Domenico, A., & Roma, P. (2024). A step forward in identifying socially desirable respondents: An integrated machine learning model considering t-scores, response time, kinematic indicators, and eye movements. *Human Behavior and Emerging Technologies, 2024*(1), 7267030. https://doi.org/10.1155/2024/7267030

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*(6), 882–888. https://doi.org/10.1037/0022-006x.51.6.882

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36*(4), 979–1016. https://doi.org/10.1111/j.0021-9029.2006.00052.x

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Meade, A. W., Pappalardo, G., Braddy, P. W., & Fleenor, J. W. (2020). Rapid response measurement: Development of a faking-resistant assessment method for personality. *Organizational Research Methods, 23*(1), 181–207. https://doi.org/10.1177/1094428118795295

Merhof, V., Böhm, C. M., & Meiser, T. (2024). Separation of traits and extreme response style in IRTree models: The role of mimicry effects for the meaningful Interpretation of estimates. *Educational and Psychological Measurement, 84*(5), 927–956. https://doi.org/10.1177/00131644231213319

Mönke, F. W., Roulin, N., Lievens, F., Bartossek, M. T., & Schäpers, P. (2024). Validity of social media assessments in personnel selection: A systematic review of the initial evidence. *European Journal of Psychological Assessment, 40*(6), 445–460. https://doi.org/10.1027/1015-5759/a000835

Moon, B., Daljeet, K. N., O'Neill, T. A., Harwood, H., Awad, W., & Beletski, L. V. (2025). Comparing the efficacy of faking warning types in preemployment personality tests: A meta-analysis. *Journal of Applied Psychology, 110*(1), 131–147. https://doi.org/10.1037/apl0001224

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*(3), 683–729. https://doi.org/10.1111/j.1744-6570.2007.00089.x

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2006). Individual differences in impression management: An exploration of the psychological processes underlying faking. *Psychology Science, 48*(3), 288–312.

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

Nie, W., Hernandez, I., Tay, L., Zhang, B., & Cao, M. (2025). A comparison of the response-pattern-based faking detection methods. *Journal of Applied Psychology, 110*(8), 1015–1035. https://doi.org/10.1037/apl0001261

Nikolaou, I., & Foti, K. (2018). Personnel selection and personality. In V. Zeigler-Hill & T. Shackelford (Eds.), *The SAGE handbook of personality and individual differences; Volume III: Applications of personality and individual differences* (pp. 458–474). Sage. https://doi.org/10.4135/9781526451248.n20

Nikolaou, I., & Katsadoraki, A. (2025). Construct validity and applicant reactions of a gamified personality assessment. *Computers in Human Behavior, 162*, 108467. https://doi.org/10.1016/j.chb.2024.108467

O'Hagan, A., Murphy, T. B., & Gormley, I. C. (2012). Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Computational Statistics & Data Analysis, 56*(12), 3843–3864. https://doi.org/10.1016/j.csda.2012.05.011

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*(4), 995–1027. https://doi.org/10.1111/j.1744-6570.2007.00099.x

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*(6), 660–679. https://doi.org/10.1037/0021-9010.81.6.660

Paek, I., & Cho, S.-J. (2015). A note on parameter estimate comparability: Across latent classes in mixture IRT modeling. *Applied Psychological Measurement, 39*(2), 135–143. https://doi.org/10.1177/0146621614549651

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

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* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-x

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Erlbaum.

Paulhus, D. L., & Trapnell, P. D. (2008). Self-presentation of personality: An agency-communion framework. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality*. Guilford.

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*(1), 158–175. https://doi.org/10.1037/a0028165

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology, 7*(4, Pt. 2), 1–18. https://doi.org/10.1037/h0025230

Plummer, M. (2017). *JAGS: Just Another Gibbs Sampler* [Computer software]. https://sourceforge.net/projects/mcmc-jags/

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics, 41*(3), 300–325. https://doi.org/10.3102/1076998616636618

Reeder, M. C., & Ryan, A. M. (2011). Methods for correcting for faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 131–150). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0087

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*(1), 3–32. https://doi.org/10.1177/01466216000241001

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*(4), 489–509. https://doi.org/10.1007/s10869-007-9038-9

Röhner, J., Schütz, A., & Ziegler, M. (2025). Faking in self-report personality scales: A qualitative analysis and taxonomy of the behaviors that constitute faking strategies. *International Journal of Selection and Assessment, 33*(1), e12513. https://doi.org/10.1111/ijsa.12513

Röhner, J., Thoss, P., & Schütz, A. (2022). Lying on the dissection Table: Anatomizing faked responses. *Behavior Research Methods, 54*(6), 2878–2904. https://doi.org/10.3758/s13428-021-01770-8

Roulin, N., Krings, F., & Binggeli, S. (2016). A dynamic model of applicant faking. *Organizational Psychology Review, 6*(2), 145–170. https://doi.org/10.1177/2041386615580875

Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*(3), 323–346. https://doi.org/10.1348/096317903769647201

Sass, R., Frick, S., Reips, U.-D., & Wetzel, E. (2020). Taking the test taker's perspective: Response process and test motivation in multidimensional forced-choice versus rating scale instruments. *Assessment, 27*(3), 572–584. https://doi.org/10.1177/1073191118762049

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*(6), 966–974. https://doi.org/10.1037/0021-9010.78.6.966

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80*(5), 607–620. https://doi.org/10.1037/0021-9010.80.5.607

Schulte, N., Holling, H., & Bürkner, P.-C. (2021). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? *Educational and Psychological Measurement, 81*(2), 262–289. https://doi.org/10.1177/0013164420934861

Schulte, N., Kaup, L., Bürkner, P.-C., & Holling, H. (2024). The fakeability of personality measurement with graded paired comparisons. *Journal of Business and Psychology, 39*(5), 1067–1084. https://doi.org/10.1007/s10869-024-09931-0

Schünemann, A. L. (2025). *On the quest for fake-proof personality assessments: Mitigating faking and socially desirable responding in low and high stakes assessment with multidimensional forced choice response formats* [Doctoral dissertation, Humboldt-Universität zu Berlin]. https://edoc.hu-berlin.de/items/179c75da-ae0e-4ac5-bf8f-6cda8fc6c446

Seitz, T., Alagöz, Ö. E. C., & Meiser, T. (2025). Disentangling qualitatively different faking strategies in high-stakes personality assessments: A mixture extension of the multidimensional nominal response model. *Educational and Psychological Measurement, 85*(6), 1237–1277. https://doi.org/10.1177/00131644251341843

Seitz, T., Spengler, M., & Meiser, T. (2023, September 29). “What if applicants fake their responses?”: Modeling faking in high-stakes in personality assessments using the multidimensional nominal response model. *PsyArXiv*. https://doi.org/10.31234/osf.io/j5mze

Seitz, T., Spengler, M., & Meiser, T. (2025). “What if applicants fake their responses?”: Modeling faking and response styles in high-stakes assessments using the multidimensional nominal response model. *Educational and Psychological Measurement, 85*(4), 747–782. https://doi.org/10.1177/00131644241307560

Seitz, T., Wetzel, E., Hilbig, B. E., & Meiser, T. (2024). Using the multidimensional nominal response model to model faking in questionnaire data: The importance of item desirability characteristics. *Behavior Research Methods, 56*(8), 8869–8896. https://doi.org/10.3758/s13428-024-02509-x

Seitz, T., & Ulitzsch, E. (2025). *Faking in high-stakes personality assessments: A response-time-based latent response mixture modeling approach* [Manuscript submitted for publication]. Department of Psychology, University of Mannheim.

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology, 65*(3), 445–493. https://doi.org/10.1111/j.1744-6570.2012.01250.x

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 19 –37). Information Age.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9*(2), 219–242. https://doi.org/10.1016/S1053-4822(99)00019-4

Speer, A. B., Delacruz, A. Y., Chawota, T., Wegmeyer, L. J., Tenbrink, A. P., Gibson, C., & Frost, C. (2025). Evaluating the impact of faking on the criterion-related validity of personality assessments. *International Journal of Selection and Assessment, 33*(1), e12518. https://doi.org/10.1111/ijsa.12518

Speer, A. B., Wegmeyer, L. J., Tenbrink, A. P., Delacruz, A. Y., Christiansen, N. D., & Salim, R. M. (2023). Comparing forced-choice and single-stimulus personality scores on a level playing field: A meta-analysis of psychometric properties and susceptibility to faking. *Journal of Applied Psychology, 108*(11), 1812–1833. https://doi.org/10.1037/apl0001099

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Sun, T., Zhang, B., Cao, M., & Drasgow, F. (2022). Faking detection improved: Adopting a Likert item response process tree model. *Organizational Research Methods, 25*(3), 490–512. https://doi.org/10.1177/10944281211002904

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408. https://doi.org/10.1007/bf02294363

Tett, R. P., Freund, K. A., Christiansen, N. D., Fox, K. E., & Coaster, J. (2012). Faking on self-report emotional intelligence and personality tests: Effects of faking opportunity, cognitive ability, and job type. *Personality and Individual Differences, 52*(2), 195–201. https://doi.org/10.1016/j.paid.2011.10.017

Tett, R. P., & Simonet, D. V. (2011). Faking in personality assessment: A "multisaturation" perspective on faking as performance. *Human Performance, 24*(4), 302–321. https://doi.org/10.1080/08959285.2011.597472

Tett, R. P., & Simonet, D. V. (2021). Applicant faking on personality tests: Good or bad and why should we care? *Personnel Assessment and Decisions, 7*(1), 6–19. https://doi.org/10.25035/pad.2021.01.002

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume 1: Models* (pp. 51–73). Chapman & Hall/CRC Press.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577. https://doi.org/10.1007/bf02295596

Ulitzsch, E., Nestler, S., Lüdtke, O., & Nagy, G. (2024). A screen-time-based mixture model for identifying and monitoring careless and insufficient effort responding in ecological momentary assessment data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000636

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika, 87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*(S1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology, 75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

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

Vargas, E. P., Carrasco-Ribelles, L. A., Marin-Morales, J., Molina, C. A., & Raya, M. A. (2024). Feasibility of virtual reality and machine learning to assess personality traits in an organizational environment. *Frontiers in Psychology, 15*, 1342018. https://doi.org/10.3389/fpsyg.2024.1342018

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197–210. https://doi.org/10.1177/00131649921969802

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology, 17*(7), 755–774. https://doi.org/10.1002/acp.914

Weijters, B., Geuens, M., & Schillewaert, N. (2010). The stability of individual response styles. *Psychological Methods, 15*(1), 96–110. https://doi.org/10.1037/a0018721

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*(2), 178–189. https://doi.org/10.1016/j.jrp.2012.10.010

Wetzel, E., & Frick, S. (2020). Comparing the validity of trait estimates from the multidimensional forced-choice format and the rating scale format. *Psychological Assessment, 32*(3), 239–253. https://doi.org/10.1037/pas0000781

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment, 33*(2), 156–170. https://doi.org/10.1037/pas0000971

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment, 23*(3), 279–291. https://doi.org/10.1177/1073191115583714

Widhiarso, W., Steyer, R., & Ravand, H. (2019). Exploring a proactive measure of making items of a personality questionnaire resistant to faking: An employee selection setting. *Personality and Individual Differences, 149*, 1–7. https://doi.org/10.1016/j.paid.2019.05.040

Wood, J. K., Anglim, J., & Horwood, S. (2022). A less evaluative measure of Big Five personality: Comparison of structure and criterion validity. *European Journal of Personality, 36*(5), 809–824. https://doi.org/10.1177/08902070211012920

Wood, J. K., Anglim, J., & Horwood, S. (2024). Less evaluative measures of personality in job applicant contexts: The effect on socially desirable responding and criterion validity. *Journal of Personality Assessment, 106*(3), 372–383. https://doi.org/10.1080/00223891.2023.2251158

Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*(1), 71–87. https://doi.org/10.1177/014662169602000107

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*(2), 168–190. https://doi.org/10.1177/1094428104263674

Ziegler, M. (2011). Applicant faking: A look into the black box. *The Industrial and Organizational Psychologist, 49*(1), 29–36.

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods, 18*(4), 679–703. https://doi.org/10.1177/1094428115574518

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). Faking: Knowns, unknowns, and points of contention. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 3–16). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0011

Behavior Research Methods (2024) 56:8869–8896
https://doi.org/10.3758/s13428-024-02509-x

ORIGINAL MANUSCRIPT

Using the multidimensional nominal response model to model faking in questionnaire data: The importance of item desirability characteristics

Timo Seitz[1] · Eunike Wetzel[2] · Benjamin E. Hilbig[2] · Thorsten Meiser[1]

Accepted: 26 August 2024 / Published online: 20 September 2024

Abstract

Faking in self-report personality questionnaires describes a deliberate response distortion aimed at presenting oneself in an overly favorable manner. Unless the influence of faking on item responses is taken into account, faking can harm multiple psychometric properties of a test. In the present article, we account for faking using an extension of the multidimensional nominal response model (MNRM), which is an item response theory (IRT) model that offers a flexible framework for modeling different kinds of response biases. Particularly, we investigated under which circumstances the MNRM can adequately adjust substantive trait scores and latent correlations for the influence of faking and examined the role of variation in the way item content is related to social desirability (i.e., item desirability characteristics) in facilitating the modeling of faking and counteracting its detrimental effects. Using a simulation, we found that the inclusion of a faking dimension in the model can overall improve the recovery of substantive trait person parameters and latent correlations between substantive traits, especially when the impact of faking in the data is high. Item desirability characteristics moderated the effect of modeling faking and were themselves associated with different levels of parameter recovery. In an empirical demonstration with *N* = 1070 test-takers, we also showed that the faking modeling approach in combination with different item desirability characteristics can prove successful in empirical questionnaire data. We end the article with a discussion of implications for psychological assessment.

Keywords Faking · Multidimensional item response theory · Item desirability · Psychological assessment · Psychometrics

When filling out a self-report personality questionnaire, test-takers have the opportunity to give overly positive self-descriptions (Paulhus, 2002). Especially when the questionnaire is part of an assessment whose results have important consequences for test-takers, a substantial proportion of test-takers can be expected to engage in faking, that is, to deliberately distort responses according to social desirability (e.g., Griffith & Converse, 2011; König et al., 2011). Unless the effect of faking is accounted for, faking can harm various psychometric properties of a test (Ziegler et al., 2011). Also, when it comes to personality assessments in actual high-stakes situations, faking can play a decisive role in decisions about hiring and promotion (e.g., Mueller-Hanson et al., 2003).

In this article, we address the response bias of faking by means of item response theory (IRT) modeling. In particular, we examine under which circumstances the multidimensional nominal response model (MNRM; Takane & de Leeuw, 1987; see Falk & Cai, 2016; Seitz et al., 2023), which offers a framework for flexibly modeling different kinds of response biases, can adequately adjust substantive trait scores and latent correlations between substantive traits for the influence of faking. We hereby focus on the role of variation in the way item content is related to social desirability (i.e., item desirability characteristics) and investigate how such variation can facilitate the modeling of faking and counteract its adverse effects.

✉ Timo Seitz
timo.seitz@uni-mannheim.de

[1] Department of Psychology, University of Mannheim, L13, 15-17 – room 515, 68161 Mannheim, Germany

[2] University of Kaiserslautern-Landau, Landau, Germany

Background: Faking in personality assessment

Faking is also known as impression management and represents the deliberate form of socially desirable responding (SDR) in Paulhus' (1984) well-known two-component model of SDR. Research has repeatedly shown that faking can have numerous effects on a test's psychometric properties (Ziegler et al., 2011). For instance, depending on whether desirable (undesirable) traits are measured, faking leads to considerably inflated (deflated) item and scale scores (e.g., Birkeland et al., 2006; Viswesvaran & Ones, 1999). A shift in item and scale scores would not be problematic for the assessment of interindividual differences if the range of possible scores was unlimited and if all test-takers shifted their scores by an equal amount. However, because self-report questionnaires often use a Likert-type rating scale with a limited number of response categories, inflated (deflated) scores are typically associated with heavily skewed score distributions and ceiling (floor) effects. Also, many studies have pointed out that test-takers differ in their propensity to fake (see Griffith & Converse, 2011). This implies that test-takers shift their scores by an unequal amount. For instance, using a randomized response technique, König et al. (2011) estimated that 32% of job applicants in the U.S. exaggerate their positive attributes in application settings whereas others do not. Likewise, when retesting job applicants under anonymous conditions (i.e., in a low-stakes context), Griffith et al. (2007) found that 30-50% of applicants had significantly elevated their scores in the preceding application (i.e., in a high-stakes context; see also Donovan et al., 2003). Such interindividual differences between test-takers imply rank-order changes and eventually alter selection decisions based on test scores (e.g., Mueller-Hanson et al., 2003). These rank-order changes can in turn have different consequences for a test's criterion-related validity, depending on how the degree of faking is correlated with the criterion variable of interest (see Komar et al., 2008). Moreover, interindividual differences in faking constitute an additional source of variance in item responses, leading to inflated intercorrelations between scales that measure desirable traits (e.g., Ellingson et al., 1999; Klehe et al., 2012; Schmit & Ryan, 1993). Faking can hence diminish construct validity in terms of a distorted discriminant validity between scales, which makes nuanced profiles of scores in a personality inventory unlikely.

Over the past decades, faking and SDR have been extensively studied by psychologists and survey methodologists. A prominent approach has been to measure SDR through designated SDR scales (see Paulhus, 2002, for an overview). These scales contain items that capture desirable behaviors hardly shown by anyone as well as undesirable behaviors that are in fact very common. Endorsing many of the former and few of the latter items would yield a high score on an SDR scale. In high-stakes assessments, SDR scales of impression management as well as related measures have been widely used to quantify faking and correct substantive trait scores for the assumed bias (Goffin & Christiansen, 2003). However, many studies have demonstrated that SDR scales are confounded with substantive trait variance and hence measure, at least to a certain degree, true personality attributes as opposed to only response bias, which makes it inappropriate to partial SDR scale scores from personality scale scores in order to achieve "pure" measures of personality (e.g., de Vries et al., 2014; McCrae & Costa, 1983; Müller & Moshagen, 2019).

Along with SDR scales and other so-called validity scales, several indirect measures have been developed to detect faking (see Goldammer et al., 2023). These include measures of response inconsistency such as person-fit indices in IRT models (e.g., LaHuis & Copeland, 2009), exploratory mixture models to identify latent faking classes (e.g., Zickar et al., 2004), and measures of extreme responding (e.g., Sun et al., 2022). However, these measures focus on the detection of faking and primarily yield an additional piece of information regarding individual test-takers. It also remains questionable how well these measures are suited to adequately adjust substantive trait scores for faking. Hence, it is appealing to have a latent variable model that directly incorporates information on the degree of faking in the estimation of model parameters and test-takers' substantive trait levels.

The multidimensional nominal response model to account for faking

To model nominal (i.e., categorial) item responses, Bock (1972) proposed an IRT model in which item responses are assumed to be influenced by a single latent dimension representing the trait of interest. Takane and de Leeuw (1987) extended this model for the case of multiple latent dimensions affecting item responses. In this multidimensional generalization, the probability of test-taker n choosing response category k out of a set of K+1 categories on item i is modeled with the following multinomial logistic function:

$$p(Y_{ni} = k|\boldsymbol{\theta}_n, \boldsymbol{\gamma}_i, \boldsymbol{\alpha}_i, \boldsymbol{S}_i) = \frac{\exp((\boldsymbol{\alpha}_i \circ \boldsymbol{s}_{ik})'\boldsymbol{\theta}_n + \gamma_{ik})}{\sum_{m=0}^{K} \exp((\boldsymbol{\alpha}_i \circ \boldsymbol{s}_{im})'\boldsymbol{\theta}_n + \gamma_{im})} \quad (1)$$

$$\text{with}\,\boldsymbol{\theta}_n = \begin{pmatrix} \theta_{n1} \\ \vdots \\ \theta_{nd} \\ \vdots \\ \theta_{nD} \end{pmatrix}, \boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{i0} & \cdots & \gamma_{ik} & \cdots & \gamma_{iK} \end{pmatrix}, \boldsymbol{\alpha}_i = \begin{pmatrix} \alpha_{i1} \\ \vdots \\ \alpha_{id} \\ \vdots \\ \alpha_{iD} \end{pmatrix},$$

$$\text{and}\ \boldsymbol{S}_i = \begin{pmatrix} s_{i10} & \cdots & s_{i1k} & \cdots & s_{i1K} \\ \vdots & & \vdots & & \vdots \\ s_{id0} & \cdots & s_{idk} & \cdots & s_{idK} \\ \vdots & & \vdots & & \vdots \\ s_{iD0} & \cdots & s_{iDk} & \cdots & s_{iDK} \end{pmatrix}.$$

Y_{ni} is a discrete random variable that reflects the response of test-taker n on item i ($Y_{ni} \in \{0, 1, \ldots, k, \ldots, K\}$), k denotes the realization of Y_{ni}, $\boldsymbol{\theta}_n$ is a D-dimensional column vector of test-taker n's levels on the D dimensions, and $\boldsymbol{\gamma}_i$ is a (K+1)-dimensional row vector of item- and category-specific intercepts. This parametrization of the MNRM (Falk & Cai, 2016; Thissen & Cai, 2016) also includes item-specific slopes α_{id} (collected in the D-dimensional column vector $\boldsymbol{\alpha}_i$) representing the relation between item i and dimension d as well as item- and category-specific scoring weights s_{idk} (collected in the D×(K+1)-dimensional matrix $\boldsymbol{S}_i$) representing the relation between dimension d and category k at item i. The symbol $\circ$ denotes the Hadamard product which links $\boldsymbol{\alpha}_i$ and $\boldsymbol{s}_{ik}$ (a column vector in matrix $\boldsymbol{S}_i$). That is, parameters pertaining to the same dimension d are multiplied before the resulting column vector is transposed and multiplied by $\boldsymbol{\theta}_n$. This leads to a sum of products $\alpha_{id}\, s_{idk}\, \theta_{nd}$ over the D dimensions. After γ_{ik} is added to this sum, the resulting term is divided by the sum of these terms for the K+1 categories to yield the probability of an item response. Hence, the MNRM falls into the class of divide-by-total IRT models (Thissen & Steinberg, 1986). For model estimation, identification constraints must be imposed (see Falk & Cai, 2016, for details). The D latent dimensions are typically assumed to be multivariate normally distributed with expectation vector $\boldsymbol{\mu} = \mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ in which all variances are fixed to 1. The intercept of the first category is usually fixed to 0 for all items.

If one has theoretical assumptions on relations between dimensions and categories, one can also specify scoring weights a priori. For latent dimensions representing substantive traits, scoring weights of items measuring the respective substantive trait are usually set to equally-spaced values (e.g., $\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix}$ in the case of a seven-point Likert scale), reflecting the assumption that higher response categories are triggered by higher substantive trait levels. Such a model is essentially a partial credit model (PCM; Masters, 1982) or a generalized partial credit model (GPCM; Muraki, 1992), depending on whether between-item equality constraints are imposed on slope parameters. Along with latent dimensions representing substantive traits, response bias dimensions can be specified. Multiple studies have used the MNRM to model response styles along with substantive traits (e.g., Bolt & Newton, 2011; Wetzel & Carstensen, 2017; see Henninger & Meiser, 2020, for an overview). Response styles are tendencies of test-takers to prefer certain response categories irrespective of item content (see Van Vaerenbergh & Thomas, 2013, for an overview). One prominent example is extreme response style (ERS), which reflects the tendency to prefer the highest or lowest category of a rating scale. Based on the definition of a particular response style, one can specify scoring weights of the respective response style dimension. For instance, in the case of a seven-point Likert scale, the scoring weight vector $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$ can be specified for ERS, reflecting the assumption that extreme rating scale categories are triggered by high ERS levels. Response styles are by definition independent of item content. Hence, the same scoring vector is usually specified for every item of the test.

To additionally account for the response bias of faking, one can add another latent dimension to the model. Because scoring weights code the relation between a dimension and a category on a particular item, scoring weights of the faking dimension can be set to values that reflect the desirability levels of response categories on a given item (Falk & Cai, 2016; see Seitz et al., 2023). As Kuncel and Tellegen (2009) demonstrated, the pattern of the relationship between response categories and social desirability differs between personality items. Hence, in contrast to substantive trait and response style dimensions, scoring weights of the faking dimension have to be specified in an item-specific way. Such a model explicitly accounts for the possibility that desirability does not increase or decrease monotonically with response categories for some items. Thus, items at which moderate levels of agreement are most desirable can be modeled, which constitutes an important extension over other recent faking models (e.g., Böckenholt, 2014; Brown & Böckenholt, 2022; Hendy et al., 2021; Leng et al., 2020; Ziegler & Bühner, 2009).

Like other psychometric models that account for response tendencies of test-takers, the presented faking model treats faking as a normally distributed latent variable. Since latent variables do not have a natural origin and scaling, the latent mean as well as the latent variance of all dimensions need to be defined for model identification. In this article, we set the latent mean to 0 and the latent variance to 1 for all dimensions. Test-takers' scores can thus be interpreted in terms of z-scores and, similar to regression analyses, intercepts represent propensities toward response categories for test-takers with mean scores on all latent dimensions. Since the fixations for model identification are arbitrary, fixing the latent faking mean to 0 does not imply that a positive versus negative faking score reflects socially desirable ("faking good") versus socially *un*desirable responding ("faking bad"). It rather reflects that a test-taker's faking degree is above versus below average in the analyzed dataset.[1] In the same vein, a faking score of 0 does not imply the absence of faking but a faking degree that corresponds to the average extent of faking in the respective sample.

Applying the presented model to a sample of bank apprentice applicants taking a Big Five personality test as part of their application, Seitz et al. (2023) provided evidence for the utility of the MNRM to model faking in a high-stakes assessment. To get scoring weights for the faking dimension, the authors had asked pilot study participants to rate each response category of each item of the personality test regarding desirability in the context of an apprenticeship in the financial industry. The model including a faking dimension with scoring weights collected in the pilot study fit the data significantly better than a model only accounting for substantive traits and response styles and improved the discriminant validity of the substantive trait scales by disinflating latent correlations. Also, comparing job applicants and job incumbents, the authors found initial evidence that the model can capture the assumed influence of faking and adjust person parameters of substantive traits in the expected direction.

Open questions

Since the study by Seitz et al. (2023) was focused on an empirical application of the model to a single high-stakes dataset and featured only a quasi-experimental validation, essential psychometric properties of the faking modeling approach are still unknown. For instance, Seitz et al. (2023) only demonstrated that the model can adjust substantive trait person parameters in the expected direction. It remained unclear if the adjustments in fact lead to more accurate estimates of the true person parameters. To answer this question, it is necessary to know the underlying population model, which is the case in simulation studies but not in applications to empirical data. Also, Seitz et al.'s (2023) empirical application mainly showed that the faking modeling approach can bring inflated latent correlations between substantive traits closer to 0. Whether it really affords more precise representations of intercorrelations between substantive traits, however, requires further research.

Along with these questions regarding the general superiority of the faking modeling approach, facilitating and limiting factors of the model's superiority have yet to be examined. For example, considering that faking is specified by setting scoring weights to desirability levels of response categories, desirability characteristics of the items used to model faking can be assumed to play a crucial role. Even though Kuncel and Tellegen (2009) found that items of regular personality tests do differ in terms of the relationship between response categories and desirability, the usual case is that higher categories are associated with higher desirability levels.[2] For instance, for 87.5% of the items in Seitz et al. (2023) and 94.5% of the items in Kuncel and Tellegen (2009), the trajectory of the relation between categories and desirability had a significantly positive linear trend. That is, personality items are in most cases constructed in a way that descriptive aspects of the trait of interest (i.e., substantive trait levels) coincide with evaluative aspects (i.e., desirability levels; Peabody, 1967). This implies that high scores can be due to a high substantive trait level, a high tendency to respond according to desirability (i.e., faking), or both, unless test-takers' faking tendency is statistically accounted for. Transferred to modeling faking by means of the MNRM, however, a situation with confounded descriptive and evaluative aspects causes high collinearity between the scoring weight vectors of the substantive trait and faking dimensions. One can assume that substantive traits and faking become increasingly hard to disentangle the more items there are with highly overlapping scoring weight vectors. In the extreme case, namely, if only one substantive trait was modeled and all items exhibited perfectly linear desirability trajectories in the direction of the substantive trait, the model would even be not identified. If descriptive and evaluative aspects were in turn not associated across items, scoring weight vectors of substantive traits and faking would not show collinearity, which arguably facilitates the modeling. Also, considering that a high faking tendency would in this case not lead to high responses on every item, high scores on a scale would be a better indication of high substantive trait levels even if faking was not accounted for. Hence, item desirability characteristics can be expected to moderate the effect of modeling faking with the MNRM.

The present research addresses the open questions regarding the MNRM approach to modeling faking by means of a simulation and an empirical study. In the simulation, it is examined if and under which conditions modeling faking along with substantive traits and response styles effectively outperforms a model without a faking dimension in terms of a) the recovery of substantive trait person parameters and b) the recovery of latent correlations between substantive traits. The empirical part in turn investigates whether the faking modeling approach in combination with different item desirability characteristics also proves successful in empirical questionnaire data.

[1] This effect is illustrated in empirical data in Supplement II.

[2] In this article, we refer to the case that items are coded such that higher categories reflect higher substantive trait levels. Also, we refer to the usual case in high-stakes personality assessments that the measured substantive traits are – on a superordinate level (i.e., independent of the content of particular items) – desirable. For substantive traits that are undesirable on a superordinate level, such as aversive personality traits, the logic reverses, that is, lower categories are generally associated with higher desirability levels.

Simulation study

Simulation design

In the present simulation, we manipulated several factors to simulate different conditions with respect to item and test construction aspects, sample size, as well as the presence of response styles and the impact of faking in the data. Irrespective of the condition, we generated data in which five substantive traits were measured by different sets of items on a seven-point Likert scale. To examine the effects of different item desirability characteristics, we compared five item compositions characterized by different levels of variety of desirability trajectories (see Fig. 1): In the first composition, all items within a substantive trait scale had a monotonically increasing desirability trajectory (i.e., highest desirability for the highest category). In the second composition, half of the items within a substantive trait scale had a desirability trajectory as in the first composition, whereas the other half had a nonmonotonically increasing desirability trajectory (i.e., desirability generally increased with higher categories but peaked at the non-extreme agreement categories and then decayed). In the third composition, two-thirds of the items within a substantive trait scale had desirability trajectories as in the second composition, whereas one-third had an inverted-U-shaped desirability trajectory (i.e., the midpoint category of the rating scale had highest desirability). In the fourth composition, three-quarters of the items within a substantive trait scale had desirability trajectories as in the third composition, whereas one-quarter had a nonmonotonically decreasing desirability trajectory (i.e., lower categories were generally associated with higher desirability, but with a peak at the non-extreme disagreement categories and a decay at the extreme disagreement category). In the fifth composition, four-fifths of the items within a substantive trait scale had desirability trajectories as in the fourth composition, whereas one-fifth had a monotonically decreasing desirability trajectory (i.e., highest desirability for the lowest category). Note that the different desirability trajectories only determined how faking manifested in item responses. Concerning substantive traits, higher response categories were always associated with higher substantive

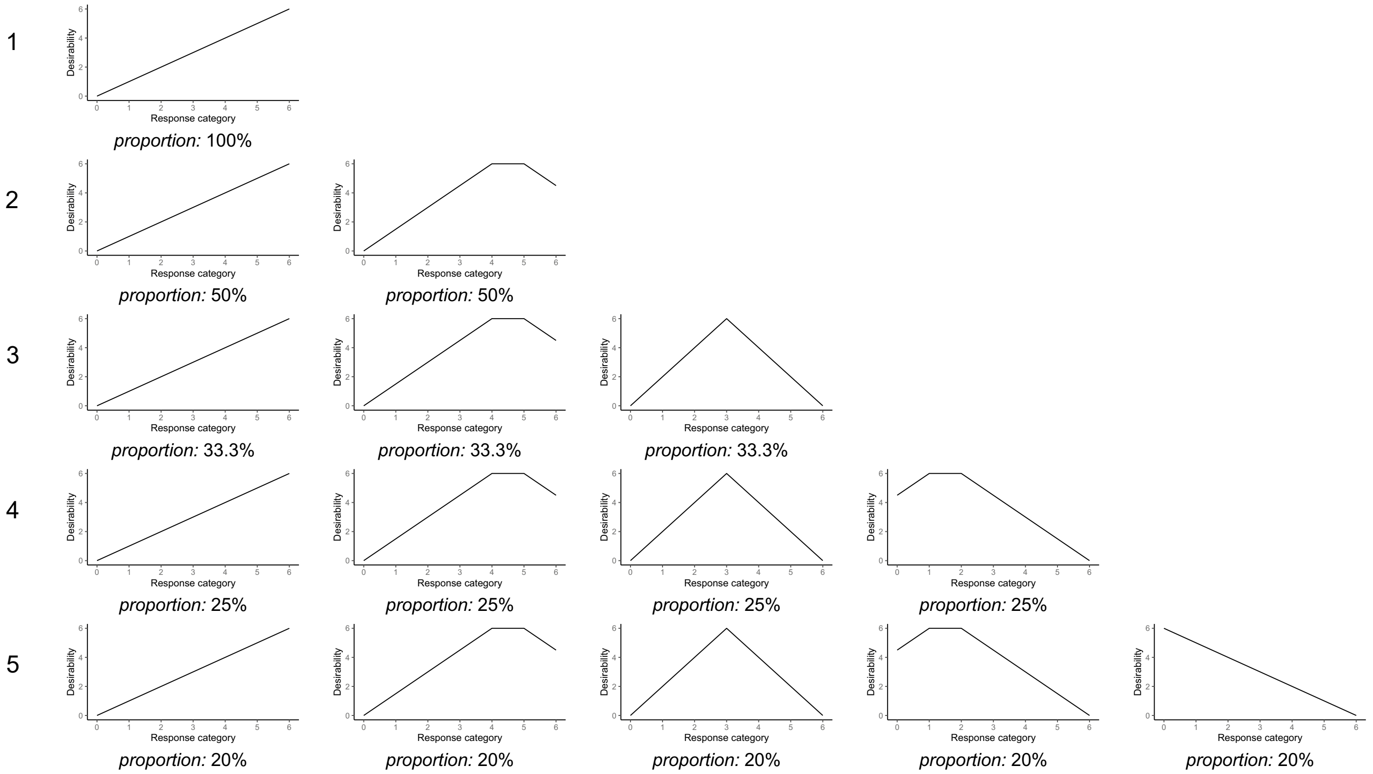


Fig. 1 Compositions of desirability trajectories. The proportions of desirability trajectories refer to the proportions within each substantive trait scale. The depicted desirability trajectories implied the following scoring weight vectors of faking: $\begin{pmatrix}0 & 1 & 2 & 3 & 4 & 5 & 6\end{pmatrix}$ for monotonically increasing trajectories; $\begin{pmatrix}0 & 1.5 & 3 & 4.5 & 6 & 6 & 4.5\end{pmatrix}$ for nonmonotonically increasing trajectories; $\begin{pmatrix}0 & 2 & 4 & 6 & 4 & 2 & 0\end{pmatrix}$ for inverted-U-shaped trajectories; $\begin{pmatrix}4.5 & 6 & 6 & 4.5 & 3 & 1.5 & 0\end{pmatrix}$ for nonmonotonically decreasing trajectories; $\begin{pmatrix}6 & 5 & 4 & 3 & 2 & 1 & 0\end{pmatrix}$ for monotonically decreasing trajectories. In conditions in which the proportions implied non-integer numbers of items, we rounded the respective proportions up or down to the next integer such that a symmetrical distribution of desirability trajectories was ensured

trait levels for all items, such that the five item compositions represented different levels of collinearity between scoring weight vectors of substantive traits and faking.

Along with item desirability characteristics, we varied the number of items per substantive trait scale (6 vs. 12) and the number of simulated test-takers (500 vs. 1000 vs. 2000). Also, we manipulated the presence of response styles (no response styles vs. ERS) and the impact of faking in the data. Considering the faking impact, we varied the extent to which the faking dimension manifested in item responses (no manifestation vs. low manifestation vs. high manifestation) to examine how this affects parameter recovery in different models.

Data generation and fitted models

To generate the data for the respective simulation conditions, we proceeded as follows (the entire simulation syntax can be found at https://osf.io/ms57p/):

1. Item-specific slopes α_{id}: Slopes of substantive trait dimensions were drawn from $U(\text{min} = 0.25, \text{max} = 0.75)$. In conditions in which ERS was present, slopes of the ERS dimension were drawn from $N(\mu = 0.25, \sigma = 0.1)$, reflecting values of a typical behavior of response styles (cf. Falk & Cai, 2016). In conditions without ERS, ERS slopes were set to 0. Regarding the impact of faking, slopes of the faking dimension were set to 0 in conditions with no faking impact, whereas faking slopes were drawn from $U(\text{min} = 0, \text{max} = 0.5)$ in low-faking impact conditions and from $U(\text{min} = 0.25, \text{max} = 0.75)$ in high-faking impact conditions. That is, faking slopes were specified to be on average as high as substantive trait slopes in conditions with a high faking impact and on average half as high as substantive trait slopes in conditions with a low faking impact.
2. Scoring weights s_{idk}: Scoring weights of substantive traits and ERS were set to values as described in the introduction of the MNRM, whereas scoring weights of ERS were linearly transformed to a range from 0 to 6 to ensure a common metric of scoring weights across dimensions (cf. Falk & Ju, 2020). Scoring weights of faking depended on the respective condition of item desirability characteristics, that is, on the composition of desirability trajectories. The respective scoring weight vectors of faking can be found in Fig. 1.
3. Item-/category-specific intercepts γ_{ik}: The intercept of the first category was fixed to 0 for all items. The remaining intercepts were generated by sampling item- and category-specific threshold values τ_{ik} from $MVN(\boldsymbol{\mu} = \overline{\boldsymbol{\tau}}, \boldsymbol{\Sigma} = \boldsymbol{T})$, where $\overline{\boldsymbol{\tau}} = (-1.5 \quad -0.9 \quad -0.3 \quad 0.3 \quad 0.9 \quad 1.5)'$ and $\boldsymbol{T} = \text{diag}(0.7 \quad 0.7 \quad 0.7 \quad 0.7 \quad 0.7 \quad 0.7)$, and transforming them to cumulative thresholds that reflect intercepts: $\gamma_{ik} = -\sum_{m=0}^{k} \tau_{im}$. These population values were chosen to generate item response distributions that could cover all response categories in the present constellation.
4. Person parameters θ_{nd}: Person parameters with a sample size depending on the respective condition were drawn from $MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$ and $\boldsymbol{\Sigma}$ was the variance-covariance matrix from Table 1. Latent variances were fixed to 1 for all dimensions. Latent covariances between substantive traits were set to values from van der Linden et al.'s (2010) meta-analysis on intercorrelations between the Big Five personality factors. ERS was set orthogonal to all substantive traits and faking. Latent covariances between faking and the five substantive traits were set to .00, .10, –.10, .30, and –.30.
5. Using the generated item and person parameters, item responses were simulated based on the multinomial logistic function in Eq. (1).
6. Steps 1 to 5 were replicated such that 100 datasets were generated per condition.

Table 1 Latent correlations between substantive traits, ERS, and faking used for data generation in the simulation

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_{ERS}	θ_{Faking}
θ_1	1						
θ_2	*.26*	1					
θ_3	*.29*	*.43*	1				
θ_4	*.36*	*.36*	*.43*	1			
θ_5	*.43*	*.21*	*.20*	*.17*	1		
θ_{ERS}	.00	.00	.00	.00	.00	1	
θ_{Faking}	.00	.10	–.10	.30	–.30	.00	1

Note. Latent correlations between substantive traits θ_1 to θ_5 are values from van der Linden et al.'s (2010) meta-analysis on intercorrelations between the Big Five (Neuroticism coded as Emotional Stability). The assignment of these correlations (printed in italics) to the ten substantive trait pairs was randomized between replications. ERS = extreme response style

All steps were carried out in the *R* environment (version 4.2.3) using the packages *MASS* (Venables & Ripley, 2002), *mirt* (Chalmers, 2012), and *SimDesign* (Chalmers & Adkins, 2020). Since research has repeatedly demonstrated the importance and stability of response styles like ERS in different assessment contexts (e.g., Bolt & Newton, 2011; LaHuis et al., 2019; Wetzel & Carstensen, 2017; Wetzel et al., 2016), a model accounting for substantive traits, ERS, and faking was compared to a model only accounting for substantive traits and ERS. These two models of interest were fitted to all 100 simulated datasets per condition.[3] For model identification, the above-described constraints were imposed. Scoring weights of the substantive trait and ERS dimensions were specified as described above. To emulate that scoring weights of the faking dimension are usually unknown in non-simulated item sets and can hence only be approximated (e.g., by pilot study ratings), we contaminated faking scoring weights used for model estimation with random noise.[4] Because of high dimensionality, models were estimated using the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010) as implemented in the *mirt* package. The MH-RM algorithm constitutes an estimation procedure that features elements from Markov chain Monte Carlo (MCMC) techniques and stochastic approximation methods and thereby converges to the maximum likelihood solution. To estimate person parameters in the high-dimensional models, maximum a-posteriori (MAP) scores were computed (see Thissen & Wainer, 2001).

Analysis

As outlined above, the simulation study should assess the performance of the faking modeling approach compared to a model not accounting for faking in recovering substantive trait person parameters and latent correlations between substantive traits under different circumstances. As the complete simulation design comprised 5 (*Item Desirability Characteristics*) × 2 (*Test Length*) × 3 (*Sample Size*) × 2 (*Presence of Response Styles*) × 3 (*Faking Impact*) = 180 conditions, we calculated effect size estimates in an analysis of variance (ANOVA) framework with the respective recovery statistic as dependent variable and the five simulation factors as well as the respective model as independent variables to evaluate the contribution of each factor and potential interactions. Since the two models of interest were fitted to the same data within a replication, we treated the factor *Model* as a repeated-measures factor. To quantify proportions of variance explained in this multifactorial mixed ANOVA, we used the R package *afex* (Singmann et al., 2023) to compute the generalized η^2 statistic (η^2_G) that provides effect size estimates that are comparable across various research designs (Olejnik & Algina, 2003). As there are no established conventions for interpreting η^2_G effect size estimates, we interpreted η^2_G values of main effects and interactions within a given ANOVA in a relative manner. Considering the large effect sizes of some main effects and interactions, we regarded η^2_G values smaller than .05 as negligible.

Concerning the recovery of substantive trait person parameters, we considered the correlation between estimated and true person parameters. In particular, the Fisher-*z*-transformed Pearson correlation between the estimated and true person parameters was computed for all five substantive traits within each replication of every condition to convert correlation coefficients into an asymptotic normal distribution. For the recovery of latent correlations, we looked at bias as well as root mean square error (RMSE). For bias, the deviation between estimated and true latent correlations was calculated for the ten substantive trait pairs *j* and then averaged within each replication of every condition:

$$\text{bias}_{rep} = \frac{1}{10} \sum_{j} (\hat{\rho}_j - \rho_j). \tag{2}$$

For RMSE, the deviation was squared and averaged across the ten substantive trait pairs before the square root was taken, which served as an indicator of estimation precision within each replication of every condition:

$$\text{RMSE}_{rep} = \sqrt{\frac{1}{10} \sum_{j} \left(\hat{\rho}_j - \rho_j\right)^2}. \tag{3}$$

Simulation results

Recovery of substantive trait person parameters

Using the above-described ANOVA framework to analyze the Fisher-*z*-transformed correlations between estimated and true person parameters of substantive traits, we found that *Model* had a main effect of $\eta^2_G = .437$, indicating considerable differences in person parameter recovery between the model ignoring faking and the model accounting for faking across conditions. The main effects of *Item Desirability*

[3] We also fitted a model only accounting for substantive traits as well as a model accounting for substantive traits and faking to all datasets. For the sake of simplicity, we provide the simulation results including these two alternative models in Supplement I. Conclusions regarding the simulation factors were identical.

[4] Random noise was generated by simulating desirability ratings of $n = 100$ hypothetical pilot study participants for each item in every replication. These ratings were based on the items' true desirability trajectories and featured a judgment of the most desirable response category for each item from every hypothetical pilot study participant. Faking scoring weights were then derived by transforming the relative frequencies of the simulated ratings per item to a range from 0 to 6.

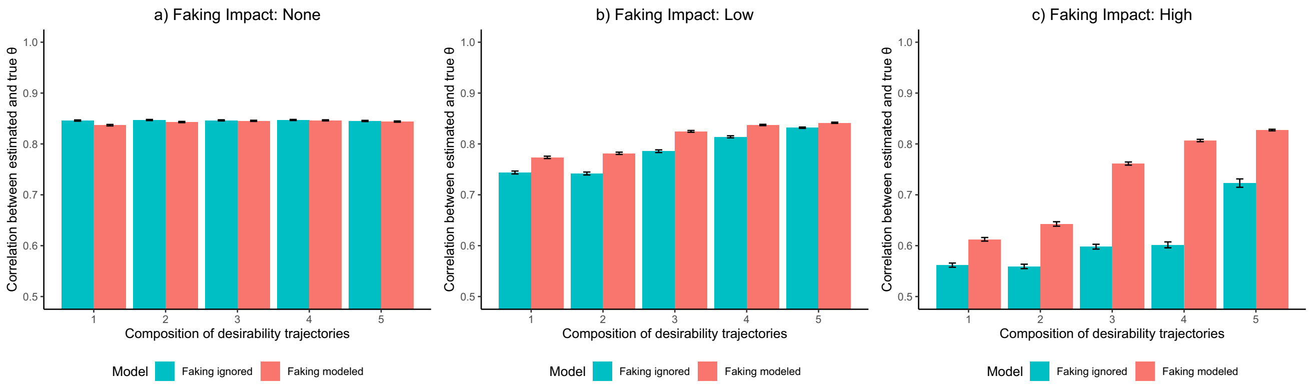


Fig. 2 Simulation study: Recovery of substantive trait person parameters. The depicted recovery of substantive trait person parameters is for the representative case of six items per substantive trait scale, a sample size of 1000, and extreme response style (ERS) being present in the data. Models ignoring faking only included dimensions for substantive traits and ERS, whereas models accounting for faking also included a faking dimension. Results are aggregated across the five substantive traits used in the simulation. Values reflect the back-transformed mean of the Fisher-*z*-transformed correlations between estimated and true person parameters across replications within a condition. Error bars represent the standard error of the mean

Characteristics (η^2_G = .565), *Test Length* (η^2_G = .678), and *Faking Impact* (η^2_G = .864) were also meaningful. In contrast, the main effects of *Sample Size* (η^2_G = .003) and *Presence of Response Styles* (η^2_G = .006) as well as all interactions including at least one of these two factors (η^2_G s < .005) were negligible. *Test Length* also did not meaningfully interact with the other factors (η^2_G s < .040), except for a two-way interaction with *Faking Impact* (η^2_G = .132). Regarding the factors *Model*, *Item Desirability Characteristics*, and *Faking Impact*, there was a pronounced three-way interaction (η^2_G = .204) that qualified the three two-way interactions between these three factors (.139 < η^2_G s < .434; effect size estimates of all main effects and interactions can be found in Table S.I.1 in Supplement I).[5]

Considering that higher-order interactions with *Test Length*, *Sample Size*, and *Presence of Response Styles* were negligible, Fig. 2 depicts the three-way interaction between *Model*, *Item Desirability Characteristics*, and *Faking Impact* for the representative case of 6 items per substantive trait scale, a sample size of 1000, and ERS being present in the data: When there was no faking in the data (see Fig. 2a), the model ignoring faking and the model accounting for faking did not differ regarding person parameter recovery, irrespective of item desirability characteristics. When the faking impact was low (see Fig. 2b), the model accounting for faking recovered person parameters better than the model ignoring faking. However, effects were rather small and almost vanished when item sets were composed of all desirability trajectory types. When the faking impact was high (see Fig. 2c), differences between the models were more pronounced, such that the model accounting for faking performed considerably better than the model ignoring faking in all compositions of desirability trajectories. Note also the main effect due to the different item desirability characteristics: Unless faking was absent in the data, person parameter recovery improved in item compositions with more variety in desirability trajectories, which was most pronounced when the faking impact was high and the model included a faking dimension.[6]

Recovery of latent correlations between substantive traits

In terms of the recovery of latent correlations between substantive traits, the above-described ANOVA framework (see Table S.I.1 for effect size estimates of all main effects and interactions) indicated that *Model* (bias: η^2_G = .257; RMSE: η^2_G = .452), *Item Desirability Characteristics* (bias: η^2_G = .567; RMSE: η^2_G = .272), and *Faking Impact* (bias: η^2_G = .594; RMSE: η^2_G = .733) had meaningful main effects on bias and RMSE. Again, there were two-way interactions (bias: .097 < η^2_G s < .579; RMSE: .057 < η^2_G s < .381) between these three factors that were qualified by a pronounced three-way interaction (bias: η^2_G = .296; RMSE: η^2_G = .165). All main effects and interactions associated with *Test Length*,

[5] We also ran an ANOVA that additionally featured the repeated-measures factor *Substantive Trait* to allow for systematic differences between the five substantive traits that had been generated to have different latent correlations with faking. The main effect of *Substantive Trait* was η^2_G= .083. All interactions including the *Substantive Trait* factor had η^2_Gs < .055. Conclusions regarding the other simulation factors were identical.

[6] As an additional analysis, we examined the recovery of person parameters of the faking dimension depending on simulation conditions. Results can be found in Table S.I.2 and Figure S.I.1 in Supplement I. Conclusions regarding the simulation factors were very similar to the ones for the recovery of substantive trait person parameters.

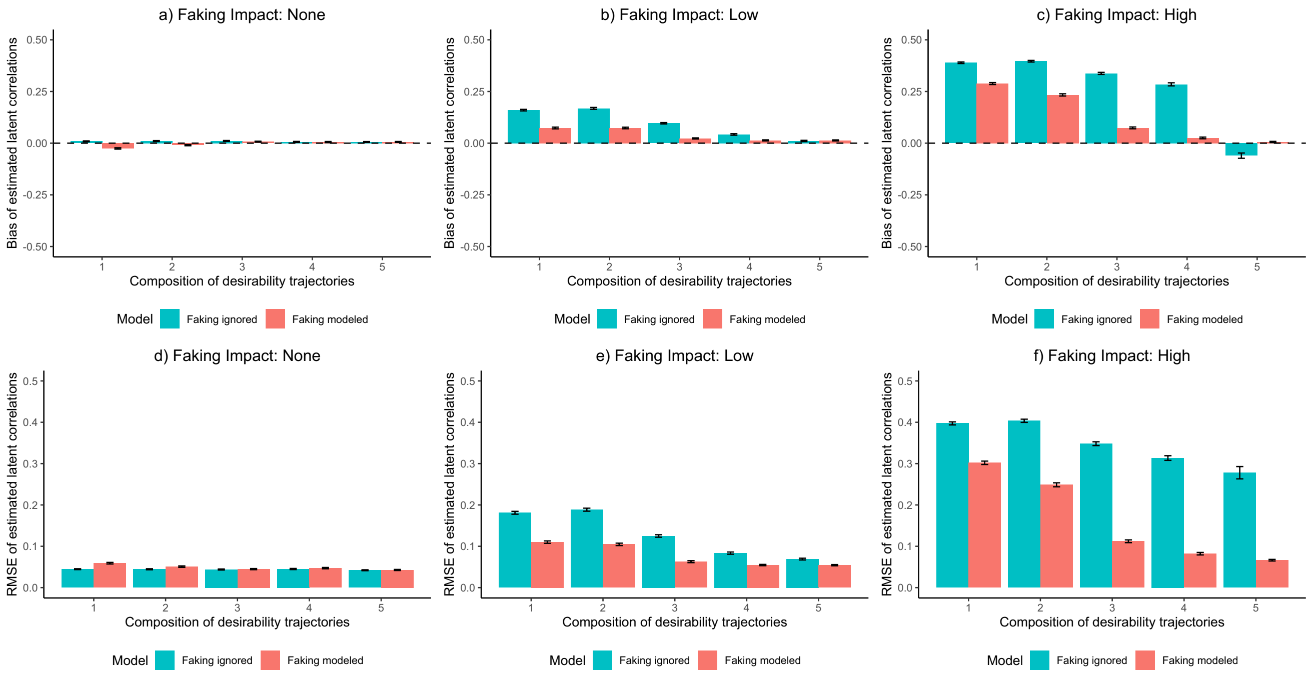


Fig. 3 Simulation study: Recovery of latent correlations between substantive traits. The depicted recovery of latent correlations between substantive traits is for the representative case of six items per substantive trait scale, a sample size of 1000, and extreme response style (ERS) being present in the data. Models ignoring faking only included dimensions for substantive traits and ERS, whereas models accounting for faking also included a faking dimension. Values reflect the mean bias (*upper panel*) and root mean square error (RMSE; *lower panel*) across replications within a condition. Error bars represent the standard error of the mean

Sample Size, and *Presence of Response Styles* were negligible (bias: all η_G^2 s < .025; RMSE: all η_G^2 s < .037).

Hence, Fig. 3 shows the three-way interaction between *Model*, *Item Desirability Characteristics*, and *Faking Impact* for the representative case of 6 items per substantive trait scale, a sample size of 1000, and ERS being present in the data: When there was no faking in the data (see Fig. 3a, d), the estimation of latent correlations between substantive traits was unbiased in both the model ignoring faking and the model accounting for faking, irrespective of item desirability characteristics. RMSE also did not differ between the different models and item desirability characteristics. However, when there was a low faking impact (see Fig. 3b, e) and desirability trajectories of items were mainly increasing, estimated latent correlations were positively biased and had larger RMSE in the model ignoring faking, whereas the model accounting for faking attenuated the bias and increased precision. Along with the effect of the model, having more variety in desirability trajectories across items also led to a reduction of bias and RMSE. The same pattern occurred when the faking impact was high (see Fig. 3c, f), but with more pronounced effects. In this case, more variety in desirability trajectories across items only led to a complete elimination of bias and a considerable reduction of RMSE in models that accounted for faking. In models that ignored faking, latent correlations were biased and imprecisely recovered in all compositions of desirability trajectories.[7]

Discussion of simulation results

The simulation results show that accounting for a faking dimension when modeling item responses that are potentially distorted by social desirability is worthwhile for estimating test-takers' substantive trait levels as well as latent correlations between substantive traits. Results indicate that the extent to which modeling faking is superior to only modeling response styles such as ERS primarily depends on the impact of faking in the data. Effects were stronger when faking explained a large proportion of variance in item responses compared to when it only explained a small proportion or when it was absent. Importantly, even when faking was not part of the data-generating process, modeling faking was not associated with a worse estimation of substantive trait person parameters. That is, modeling faking

[7] As an additional analysis, we examined the recovery of latent correlations between the faking dimension and substantive traits depending on simulation conditions. Results can be found in Table S.I.2 and Figure S.I.2 in Supplement I. Conclusions regarding the simulation factors were very similar to the ones for the recovery of latent correlations between substantive traits.

with the MNRM does not erroneously attribute substantive trait variance to a faking dimension, which is a major limitation of SDR scales (e.g., de Vries et al., 2014; McCrae & Costa, 1983; Müller & Moshagen, 2019) and has been observed in the context of response styles (e.g., Merhof et al., 2023). Note that the findings were independent of the number of items per substantive trait scale, the number of test-takers, and the presence of ERS in the data. However, item desirability characteristics played an important role such that they moderated the effect of modeling faking and led themselves to different levels of recovery of substantive trait person parameters.

Regarding the recovery of latent correlations between substantive traits, the simulation yields a similar conclusion. Modeling faking led to less biased and more precise intercorrelations between substantive traits when faking was present in the data, particularly when the impact of faking was high. This indicates that modeling faking with the MNRM can indeed debias inflated correlations between substantive traits (e.g., Ellingson et al., 1999; Klehe et al., 2012; Schmit & Ryan, 1993) and thus facilitate nuanced test-taker profiles within a personality inventory. As for the recovery of substantive trait person parameters, the simulation findings were independent of test length, sample size, and presence of response styles, whereas having more variety in desirability trajectories across items interacted with the effect of modeling faking and could per se improve the recovery of latent correlations between substantive traits.

Empirical demonstration

The simulation study had the purpose of investigating the potential of modeling faking with the MNRM when the data-generating process and true parameter values are known. To examine whether the faking modeling approach in combination with different item desirability characteristics also proves successful in empirical data, we collected questionnaire data using an experimental faking manipulation and a special set of items. To emulate that desirability depends on the social situation (e.g., Ellingson & McFarland, 2011; Kuncel & Tellegen, 2009) and that faking is inherent to the assessment context at hand, we used a specific social context for item responding, namely a hypothetical application for a leadership position in the industry.

Development of items with different desirability characteristics

As noted by Peabody (1967) and other scholars (e.g., Bäckström et al., 2009; Wood et al., 2022), most personality items are constructed in a way that descriptive and evaluative aspects are confounded, that is, high rating scale categories are associated with both high substantive trait levels and high desirability levels. Hence, to examine the effects of different item desirability characteristics in empirical questionnaire data, we adapted items from a widely-used personality test, the German version of the *Big Five Inventory 2* (BFI-2; Danner et al., 2016, 2019), such that they should still measure the Big Five but deconfound substantive trait levels and desirability levels. That is, we modified the BFI-2 items to create more items with nonmonotonically increasing, inverted-U-shaped, and decreasing desirability trajectories. Note that this approach is different from the approach followed by Bäckström et al. (2009, 2023) and Wood et al. (2022, 2023), who merely aimed at reducing evaluative item content to counteract SDR instead of creating more variety in desirability trajectories. We then piloted the original and modified items to obtain empirical desirability ratings.

Steps of item modification

Before modifying items, we conducted a brief job demand analysis for a leadership position in the industry to get a better understanding of what is considered desirable and undesirable in this particular social context. We therefore looked for articles in the leadership literature portraying the role of different personality attributes (e.g., Ames & Flynn, 2007; Baron et al., 2000; De Hoogh et al., 2005; Hogan & Kaiser, 2005; Judge et al., 2002; Kaiser et al., 2015) and surveyed ten persons currently holding a leadership position. Based on the results of the job demand analysis, we then reworded items from the BFI-2 with the aim of creating modified items where higher rating scale categories are not monotonically related to higher desirability levels but where higher categories are still related to higher levels of the to-be-measured Big Five trait.[8] Following certain strategies of item modification (see Supplement II for details), we created 104 modified items. In the next step, the modified items were reviewed regarding their fit to the construct definitions of the Big Five (McCrae & Costa, 1987) as well as regarding aspects like ambiguity and item length. This led to an exclusion of 39 modified items. The 60 original BFI-2 items as well as the 65 retained modified items can be found in Table A1 in the Appendix.

[8] Note that this logic reverses for negatively-keyed items. These items are phrased such that lower categories represent higher levels of the substantive trait. Hence, these items as well as their desirability trajectories need to be recoded.

Pilot study

We then piloted the original and modified items to obtain empirical desirability ratings for the context of an application for a leadership position in the industry. Because we carried out the modification of items in two waves, we ran two piloting rounds to obtain desirability ratings for all 125 items. The study procedure and the population for participant sampling were, however, identical between the two piloting rounds. Forty-one modified items as well as the 60 BFI-2 items were piloted in the first round, the 24 remaining modified items in the second round.

Procedure We ran the pilot study on the online data collection platform *SoSci Survey* (https://www.soscisurvey.de/). After giving informed consent and completing demographic measures, participants were asked to take the perspective of a person who is currently applying for a leadership position in the industry. We then familiarized participants with typical tasks of personnel in leadership positions before telling them that the application process would feature a questionnaire on personal attitudes and behaviors. Next, we informed them about their task: For every statement (i.e., item), they were instructed to judge which of seven graded agreement levels (i.e., response categories; 1 = *very low agreement* to 7 = *very high agreement*) is most desirable in the given context. Items were presented in a random order and on separate pages. After half of the items, participants could take a self-paced break. The exact instructions, data, as well as analysis code of the pilot study can be found at https://osf.io/ms57p/.

Sample To obtain desirability ratings from people who could potentially apply for a leadership position, we allowed participation in the pilot study only if participants were at least 18 years old and already had work experience. Because the items were created in German, participants also needed to speak German fluently to be eligible for participation. We excluded participants from the analyses if they had failed at least one instructed-response item (e.g., “Please click here on scale point 3”), if they indicated that their data shall not be used, or if their median item response time was less than 50% of the median item response time across all participants (cf. Gummer et al., 2021; Leiner, 2019). The participant samples of the two piloting rounds had a similar distribution of age, gender, work experience, and leadership experience. The sample of the first round (N = 152) had a mean age of M = 28.43 years (SD = 13.06, range = [18–71]), with 71.1% being female (28.9% male). The mean work experience was M = 7.88 years (SD = 11.77, range = [1–41]), with the majority (85.5%) never having held a leadership position. The mean age in the sample of the second round (N = 196) was M = 26.01 years (SD = 11.20, range = [18–65]) and 73.0% were female (27.0% male). Participants in this sample had a mean work experience of M = 5.80 years (SD = 9.61, range = [1–49]) and 88.3% had never held a leadership position. The two samples did not differ significantly (α = .05) on any of the four demographic variables ($|t|$s < 1.87, ps $>$ .062; χ^2s < 0.36, ps $>$ .552).

Results After recoding desirability ratings for negatively-keyed items, we calculated Pearson’s X^2 statistic for all desirability trajectory types t and each item i to classify the piloted items into the five desirability trajectory types that were also used in the simulation study:

$$X_{ti}^2 = \sum_{k=0}^{6} \frac{\left(O_{ki} - E_{tk}\right)^2}{E_{tk}}. \tag{4}$$

E_{tk} denotes the expected frequency of desirability ratings of response category k under desirability trajectory type t. These values were derived from the prototypical desirability trajectories shown in Fig. 1 and the number of participants giving desirability ratings for the respective item. O_{ki} represents the observed frequency of desirability ratings of category k on item i. We then classified the items into the five desirability trajectory types based on the minimal X^2 value that resulted for a given item. According to this classification, 38 (30.4%) of the 125 original and modified items had monotonically increasing desirability trajectories, 29 (23.2%) exhibited nonmonotonically increasing desirability trajectories, 32 (25.6%) inverted-U-shaped desirability trajectories, 19 (15.2%) nonmonotonically decreasing desirability trajectories, and 7 (5.6%) monotonically decreasing desirability trajectories. The classification of each item can be found in Table A1. Figure 4 shows histograms of desirability ratings for five exemplary items.

Main study: Collecting item responses under low-stakes and high-stakes conditions

Design

After modifying the BFI-2 items and piloting them together with the original items concerning their desirability trajectories, we gave the whole item set to another sample of participants and instructed them to respond to the items under two conditions, namely an honest condition as well as a hypothetical application condition. The honest condition served as a low-stakes (LS) condition in which we asked participants to respond as honestly as possible. The hypothetical application condition served as an experimental high-stakes (HS) condition. In this condition, participants were instructed to respond as if they were applying for a leadership position in the industry. Also, they received a financial incentive to adapt their responses to meet the requirements

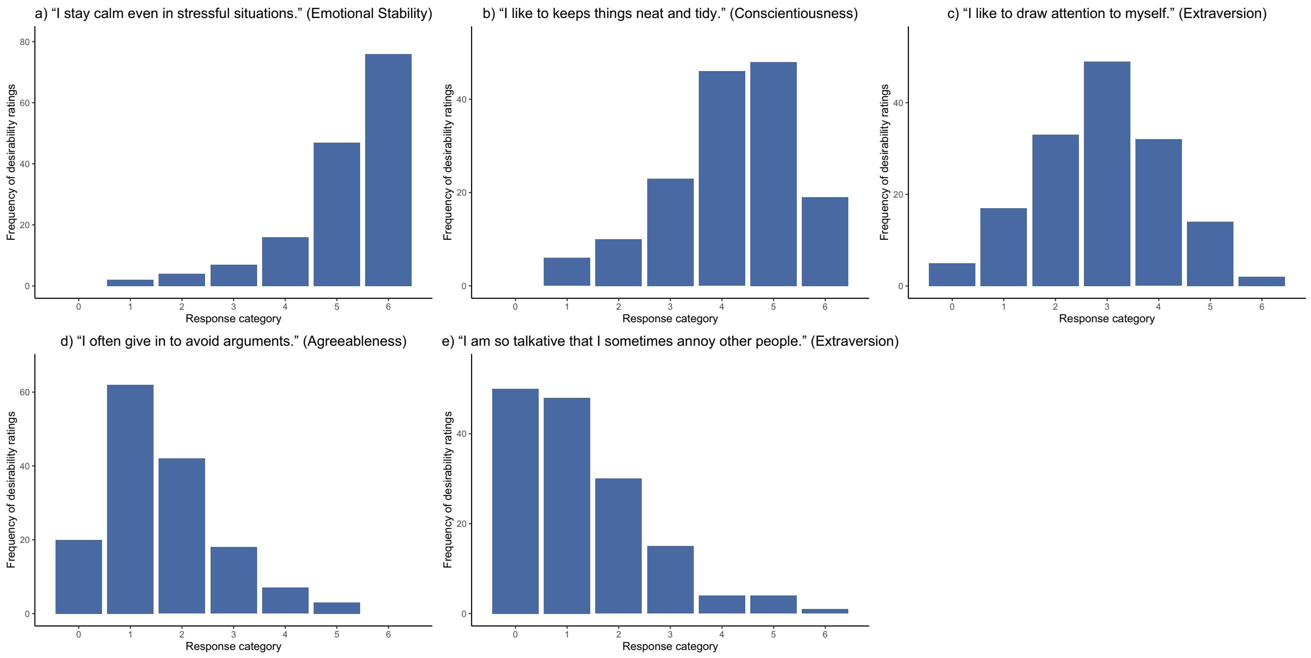


Fig. 4 Empirical demonstration: Histograms of desirability ratings for five exemplary items. N = 152 for all of the five exemplary items. Classifications of desirability trajectories: **a**) monotonically increasing, **b**) nonmonotonically increasing, **c**) inverted-U-shaped, **d**) nonmonotonically decreasing, **e**) monotonically decreasing

of the advertised job. To control potential carry-over effects between the two conditions, we randomized the order of conditions between participants.

Procedure

SoSci Survey served as the online platform for data collection. At the beginning, participants were asked to give informed consent and complete demographic measures. Subsequently, they read the instructions of the first condition and responded to the original and modified items before reading the instructions of the second condition and responding again to all items. In the LS condition, we emphasized that there were no right or wrong answers and that the data were kept strictly confidential. In the HS condition, we asked participants to take the perspective of a person who is currently applying for a leadership position at a fictitious company in the industry. Therefore, participants saw a fictitious job advertisement for the vacant leadership position in which the company communicated tasks of their leadership personnel as well as their expectations about the personality of applicants. We then told participants that, in order to identify applicants who fit ideally to the vacant position, a questionnaire about personal attitudes and behaviors would be part of the application process. Subsequently, we instructed participants to respond to the items as if they were in the described application context. However, as is usually the case in high-stakes assessments, we asked them to respond based on their actual attitudes and behaviors, but at the same time try to get the vacant position. To create actual stakes for participants, we told them that the 10% of participants matching the personality profile from the job advertisement best would receive the double amount of the standard compensation for participation.[9] The exact instructions from the two conditions can be found at https://osf.io/ms57p/. In both conditions, responses were given on a seven-point Likert scale (1 = *very low agreement* to 7 = *very high agreement*). In both conditions, items were presented in a randomized order and on separate pages. Participants had the opportunity to take a self-paced break after half of the items had been presented in each condition. However, before participants could respond to the items in a condition, they had to pass a diligence check item in which they were queried about how they were supposed to respond according to the instructions of the respective condition. After completing both conditions, participants were debriefed and thanked for participating. The completion of the entire study took approximately 30 minutes.

[9] After data collection, we calculated the mean absolute deviation between item responses and modes of desirability ratings from the pilot study for each participant of the main study across all 125 items. The 10% of participants with the smallest mean absolute deviation received the bonus compensation.

Sample

The sample of participants was collected via *Bilendi*, a European access panel service provider. Participants successfully completing the study received compensation worth 5 euros. As in the pilot study, we allowed participation only if participants were at least 18 years of age, already had work experience, and spoke German fluently. To ensure good data quality, we a priori implemented several quality checks in a way that participants failing at least one of these quality checks were immediately screened out and could not finish the data collection. In particular, we used the following quality checks (cf. Gummer et al., 2021; Leiner, 2019): two instructed-response items, the above-mentioned diligence check items querying participants about the preceding instructions, a self-report diligence check where participants could indicate that their data shall not be used, a longest-string analysis where participants were screened out if they provided at least ten identical responses to consecutive items in the LS condition, as well as a response time criterion where participants were screened out if their median item response time was less than 50% of the median item response time across the participants who had previously passed all quality checks. Because participants failing at least one of these quality checks were screened out during data collection, their data were not available for analysis and no post-hoc exclusion of participants needed to be made. The sample of participants passing all quality checks comprised $N = 1070$ subjects. Within this sample, the mean age was M = 36.78 years (SD = 13.06, range = [18–65]) and 54.0% were female (46.0% male). Regarding work and leadership experience, participants reported a mean work experience of M = 14.50 years (SD = 12.76, range = [1–53]) and 65.6% had never held a leadership position.

Results of the empirical demonstration

Since the faking manipulation was operationalized within participants, it was possible to analyze item responses from one sample of test-takers in both an LS and HS condition. Considering the instructions as well as the financial incentive to distort responses in the HS condition, parameter estimates from models fitted to HS condition data could be expected to be systemically biased by faking. In contrast, given that participants were instructed to respond as honestly as possible in the LS condition and had no obvious motivation to present themselves in an overly favorable manner, parameter estimates from models fitted to LS condition data should not be systematically influenced by faking but represent approximations of true parameter values. This allowed us to compare the parameter estimates from the model ignoring faking and the model accounting for faking (both fitted to the HS condition data) regarding the question of which model represents the approximated true parameter values (i.e., estimates from the LS condition data) better.

Composition of item sets with different desirability trajectories

To be able to examine the effects of different item desirability characteristics, we composed five item sets mirroring the five compositions of desirability trajectories from the simulation study (see Fig. 1). Each item set consisted of 30 items, with 6 items per Big Five trait. Descriptively, the desirability trajectory type of an item was strongly related to the mean shift of item responses between the LS and HS condition (polyserial correlation of .84, p < .001): Items with monotonically increasing desirability trajectories had an item mean that was on average 0.74 scale points higher in the HS than in the LS condition, whereas items with nonmonotonically increasing desirability trajectories yielded a mean shift of 0.56, items with inverted-U-shaped desirability trajectories a mean shift of 0.31, items with nonmonotonically decreasing desirability trajectories a mean shift of –0.27, and items with monotonically decreasing desirability trajectories a mean shift of –0.58. To compose the five item sets in the proportions of desirability trajectories as indicated in Fig. 1, we selected the items that measured the underlying Big Five trait best. Therefore, we fitted a model with all 125 items to the LS condition data, in which only substantive traits and ERS were accounted for (Bolt & Newton, 2011; Wetzel & Carstensen, 2017). In this model, we considered the estimated item slopes to select the items with highest discrimination concerning the underlying Big Five trait in a dataset where item responses should not be systematically influenced by faking. To ensure that the meaning of the substantive traits as measured in the BFI-2 did not fundamentally change when adding the modified items, we fixed the slopes of the BFI-2 items to the estimated values from a corresponding model in which only the 60 BFI-2 items were modeled in the LS condition data. The five item sets that were composed based on this procedure as well as reliability estimates and convergent validities with the original BFI-2 can be found in Table S.II.1 in Supplement II.

Fitted models

Within each item composition, we fitted different models to the HS condition data. We used the MH-RM algorithm implemented in the R package *mirt*, imposed model identification constraints as described above, and estimated person parameters via MAP scores. In particular, we fitted a model only accounting for substantive traits, a model accounting for substantive traits and ERS, as well as a model additionally accounting for faking. As in the simulation, we linearly transformed the scoring weights of the ERS and faking dimensions to a range from 0 to 6 for a common metric of scoring weights across all dimensions (cf. Falk & Ju, 2020). Such linear transformations of scoring weights do not affect the estimation of person parameters, latent correlations, and model fit. Scoring weights of the faking dimension were based on the relative frequencies of desirability ratings from the pilot study, which can be found in Table S.II.1.

All models converged within 718 MH-RM iterations. Table 2 contains absolute (Cai & Monroe, 2013, 2014; Maydeu-Olivares & Joe, 2014) and relative model fit measures for the fitted models. In all compositions of desirability trajectories, fit measures consistently indicated that modeling ERS improved model fit compared to only modeling substantive traits. Crucially, adding a faking dimension improved model fit further in all item compositions, showing that the faking modeling approach could explain incremental variance in item responses over and above response styles.

Correlations between low-stakes and high-stakes substantive trait person parameters

Assuming that person parameters of substantive traits in the LS condition were not systematically biased by faking, we looked at correlations of substantive trait person parameters between the LS and HS condition to examine if the model including a faking dimension could recover the substantive trait person parameters from the LS condition better than a model not accounting for faking. Modeling ERS yielded a significantly better model fit than not accounting for response styles in all item compositions also in the LS condition data (χ^2s(35) > 3390.4, *p*s < .001). Hence, we computed correlations of substantive trait person parameters from the model accounting for substantive traits and ERS in the LS condition a) with person parameters from the corresponding model in the HS condition and b) with person parameters from the model additionally accounting for faking in the HS condition. We compared these two correlations for all Big Five traits in the five item compositions.

Table 2 Empirical demonstration: Model fit measures of fitted models

	Absolute fit measures				Relative fit measures			
Dimensions modeled	C_2 (*df*), *p*-value	RMSEA	CFI	TLI	Log-likelihood	AIC	BIC	LR test
Composition of desirability trajectories 1:								
θs	3059.9 (395), *p* < .001	.079	.947	.952	–40815.2	82070.5	83165.1	
θs/ERS	2687.2 (360), *p* < .001	.078	.949	.958	–38850.4	78210.9	79479.6	$\chi^2(35) = 3929.6, p < .001$
θs/ERS/Faking	**1537.3 (324), *p* < .001**	**.059**	**.971**	**.978**	**–38459.1**	**77500.2**	**78948.0**	$\boldsymbol{\chi^2(65) = 782.7, p < .001}$
Composition of desirability trajectories 2:								
θs	2444.2 (395), *p* < .001	.070	.942	.947	–44691.3	89822.5	90917.1	
θs/ERS	2166.1 (360), *p* < .001	.069	.944	.954	–42857.0	86224.0	87492.7	$\chi^2(35) = 3668.5, p < .001$
θs/ERS/Faking	**1684.1 (324), *p* < .001**	**.063**	**.953**	**.965**	**–42688.9**	**85959.8**	**87407.7**	$\boldsymbol{\chi^2(36) = 336.1, p < .001}$
Composition of desirability trajectories 3:								
θs	3186.7 (395), *p* < .001	.081	.896	.906	–47609.5	95659.1	96753.7	
θs/ERS	2882.4 (360), *p* < .001	.081	.897	.915	–45723.5	91956.9	93225.7	$\chi^2(35) = 3772.2, p < .001$
θs/ERS/Faking	**2100.9 (324), *p* < .001**	**.072**	**.919**	**.940**	**–45493.3**	**91568.5**	**93016.4**	$\boldsymbol{\chi^2(36) = 460.4, p < .001}$
Composition of desirability trajectories 4:								
θs	3900.6 (395), *p* < .001	.091	.787	.807	–51303.0	103046.0	104140.6	
θs/ERS	2575.2 (360), *p* < .001	.076	.853	.878	–49235.7	98981.4	100250.1	$\chi^2(35) = 4134.6, p < .001$
θs/ERS/Faking	**1614.8 (324), *p* < .001**	**.061**	**.905**	**.929**	**–48897.5**	**98377.0**	**99824.8**	$\boldsymbol{\chi^2(36) = 676.4, p < .001}$
Composition of desirability trajectories 5:								
θs	3716.8 (395), *p* < .001	.089	.797	.816	–52092.2	104624.5	105719.1	
θs/ERS	2275.6 (360), *p* < .001	.071	.872	.894	–49828.4	100166.7	101435.5	$\chi^2(35) = 4527.7, p < .001$
θs/ERS/Faking	**1486.3 (324), *p* < .001**	**.058**	**.914**	**.936**	**–49496.2**	**99574.5**	**101022.3**	$\boldsymbol{\chi^2(36) = 664.3, p < .001}$

Note. N = 1070. The five compositions of desirability trajectories correspond to those displayed in Fig. 1. Models were fitted to data from the high-stakes (HS) condition. C_2 = limited information fit statistic C_2; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index; AIC = Akaike information criterion; BIC = Bayesian information criterion; LR test = likelihood-ratio test (here: hierarchical comparison of nested models); ERS = extreme response style. The best-fitting model within each item composition is printed in bold

For 22 of the 25 comparisons, person parameters from the LS condition were descriptively more strongly associated with person parameters from the HS condition when faking was modeled in the HS condition data than when faking was not modeled. The difference in correlations reached significance ($\alpha = .05$, one-tailed tests) for 19 of the 25 comparisons. All correlations and significance tests can be found in Table S.II.2 in Supplement II. Figure 5 shows the mean correlations between the LS and HS condition across the Big Five traits for the model ignoring faking and the model accounting for faking in the five item compositions. Along with the overall higher correlations for the model accounting for faking, it is noticeable that differences in correlations were more pronounced in item compositions that predominantly consisted of items with increasing desirability trajectories as compared to item compositions that also contained items with decreasing desirability trajectories. Moreover, the latter item compositions yielded generally higher correlations of person parameters than the former item compositions. This pattern of results mirrors the result pattern from the simulation study for the case of a low faking impact (see Fig. 2).

Latent correlations between substantive traits

Under the assumption that item responses in the LS condition were not systematically affected by faking, the estimated latent correlations between substantive traits in the LS condition should serve as unbiased approximations of the intercorrelations between the substantive traits as measured in the present items. To examine the effects of modeling faking as well as the different item desirability characteristics on the estimation of latent correlations between substantive traits in the HS condition, we calculated the mean bias and RMSE of the estimated latent correlations between the Big Five with respect to the corresponding latent correlations from the model accounting for substantive traits and ERS in the LS condition. Results are displayed in Fig. 6.

Regarding bias, the model accounting for faking reduced the bias of latent correlations compared to the model ignoring faking in item compositions that contained items with increasing desirability trajectories. When there were also items with inverted-U-shaped desirability trajectories, the model accounting for faking induced a slight negative bias, whereas latent correlations in the model ignoring faking were still considerably positively biased. When decreasing desirability trajectories were present, latent correlations in both models were almost unbiased. This pattern generally mirrors the findings from the simulation (see Fig. 3). Concerning RMSE, latent correlations in the model accounting for faking had considerably smaller RMSE than latent correlations in the model ignoring faking when items only had increasing desirability trajectories, which is in line with the simulation findings. However, unlike in the simulation, more variety in desirability trajectories across items was not associated with smaller RMSE in the model accounting for faking. However, when item compositions also contained inverted-U-shaped and/or decreasing desirability trajectories, RMSE in the model accounting for faking was still smaller than in the model ignoring faking.

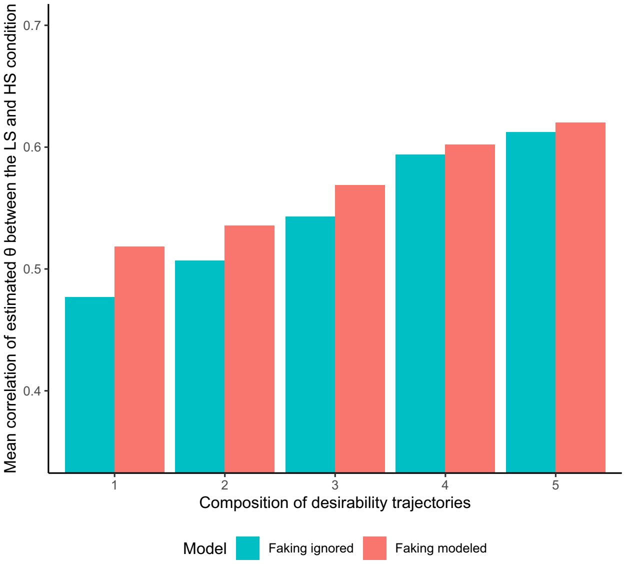


Fig. 5 Empirical demonstration: Mean correlations of substantive trait person parameters between the low-stakes and high-stakes condition. The depicted mean correlations are aggregated across the Big Five and reflect the back-transformed means of the Fisher-z-transformed correlations of substantive trait person parameters between the low-stakes (LS) and high-stakes (HS) condition. To the LS condition data, a model ignoring faking was fitted, whereas a model ignoring faking and a model accounting for faking were fitted to the HS condition data. The five compositions of desirability trajectories correspond to those displayed in Fig. 1. Models ignoring faking only included dimensions for substantive traits and ERS, whereas models accounting for faking also included a faking dimension

Discussion of results of the empirical demonstration

The purpose of the empirical demonstration was to investigate if the MNRM approach to modeling faking in combination with different item desirability characteristics also proves successful in empirical questionnaire data. Replicating the findings from Seitz et al. (2023), the results show that modeling faking can improve model fit over and above modeling response styles in experimental high-stakes assessment data.

Importantly, the results also demonstrate that modeling faking can overall increase the extent to which low-stakes substantive trait person parameters, which serve as benchmarks of trait assessment not influenced by deliberate faking, are recovered in high-stakes data. Note that this was found although the model in the LS condition and the

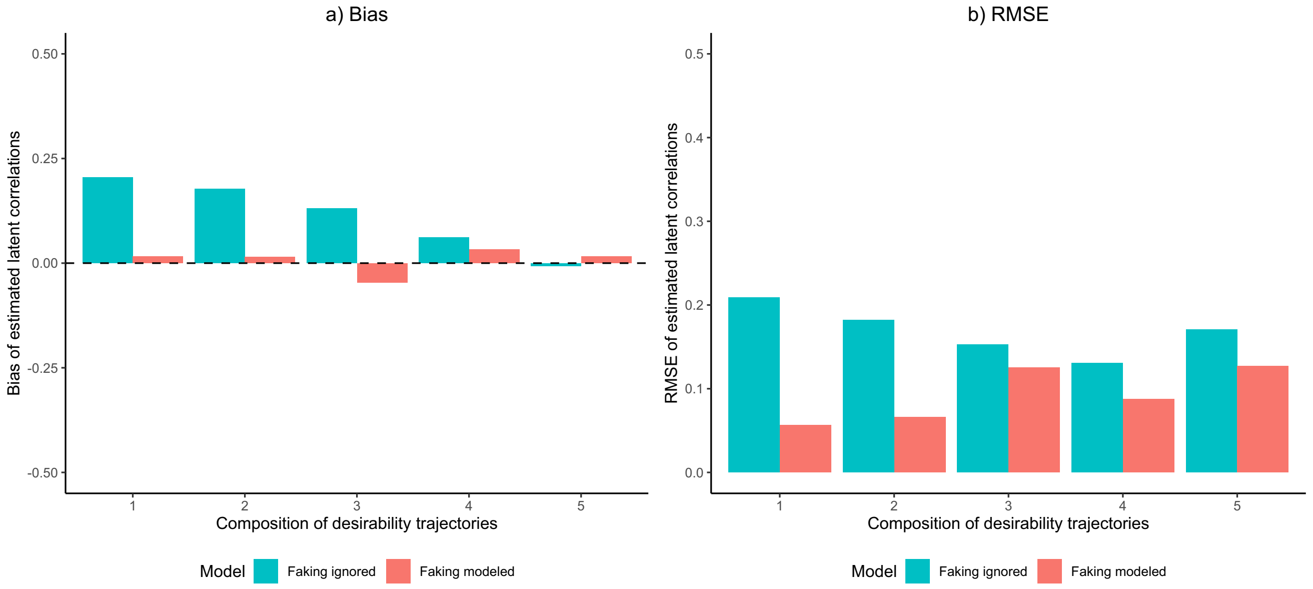


Fig. 6 Empirical demonstration: Bias and root mean square error (RMSE) of estimated latent correlations between substantive traits. The depicted values refer to the mean bias and root mean square error (RMSE) of estimated latent correlations between substantive traits in the high-stakes (HS) condition with respect to the latent correlations from a model ignoring faking in the low-stakes (LS) condition. To the LS condition data, a model ignoring faking was fitted, whereas a model ignoring faking and a model accounting for faking were fitted to the HS condition data. The five compositions of desirability trajectories correspond to those displayed in Fig. 1. Models ignoring faking only included dimensions for substantive traits and ERS, whereas models accounting for faking also included a faking dimension

model ignoring faking in the HS condition had the same dimensional structure (namely, five substantive traits and ERS), whereas the model accounting for faking additionally included a faking dimension. Moreover, effects were moderated by item desirability characteristics such that the effect of modeling faking was most pronounced in item compositions with predominantly increasing desirability trajectories. In item compositions also containing decreasing desirability trajectories, the effect of the modeling was practically negligible. Additionally, item compositions with more variety in desirability trajectories were associated with generally better recovery of low-stakes person parameters in high-stakes data, irrespective of whether faking was modeled.

Regarding latent correlations between substantive traits, the results also replicate Seitz et al.'s (2023) findings to the effect that adding a faking dimension to the model can debias inflated latent correlations between substantive traits and increase the precision of estimation. Like for substantive trait person parameters, the results indicate that more variety in desirability trajectories across items can have a debiasing effect regardless of whether or not the model contains a faking dimension. Estimation precision, however, did not consistently improve with more variety in desirability trajectories across items.

Overall, the findings of the empirical demonstration are aligned with the simulation results. Concerning both the pattern of results and the relatively small effect sizes concerning the estimation of substantive trait person parameters, the findings are especially in line with the case of a low faking impact in the data (see Figs. 2 and 3). This constitutes a plausible finding considering the experimental nature of the present HS condition, in which all participants received the same explicit instruction to respond based on their actual attitudes and behaviors but try to get the vacant position at the same time. It is possible that the strong situational cues in the experimental setting, in which no strong differences in the motivation to adhere to the instructions can be expected, led to restricted variation in the degree of faking between participants (cf. Birkeland et al., 2006; McFarland & Ryan, 2000), implying a relatively low impact of a latent faking dimension. Psychometrically, this is reflected in the estimated slopes of the faking dimension, which were on average notably smaller ($\overline{\alpha}_{\cdot\text{Faking}} = 0.12$) than the estimated slopes of the substantive trait dimensions ($\overline{\alpha}_{\cdot\theta\text{s}} = 0.80$).

General discussion

In the present research, we applied IRT modeling to account for the response bias of faking. Using a simulation and an empirical demonstration, we investigated under which circumstances the MNRM approach to modeling faking can adequately adjust substantive trait scores and latent correlations between substantive traits for the influence of faking. In particular, we were interested in how different item desirability characteristics can facilitate the modeling of faking and counteract its detrimental effects.

Utility of modeling faking

As outlined in the introduction, the faking modeling approach of this article entails an item-dependent specification of how response categories are related to social desirability and thus allows for a confirmatory modeling of faking. In contrast, approaches that aim to account for faking in a data-driven manner (e.g., exploratory mixture models with latent faking classes; Zickar et al., 2004) do not directly justify the assertion that it is faking and not an unknown combination of other response biases that has been accounted for. Furthermore, the model allows for curvilinear relationships between response categories and social desirability (cf. Kuncel & Tellegen, 2009) because scoring weights of faking are specified in an item- and category-specific manner. This constitutes an important advantage over other recent latent variable models of faking (e.g., Böckenholt, 2014; Brown & Böckenholt, 2022; Hendy et al., 2021; Leng et al., 2020; Ziegler & Bühner, 2009), which do not explicitly account for item-specific trajectories of social desirability over response categories and hence neglect this relevant information in item responses.

Overall, we found that the adjustments of substantive trait person parameters afforded by the MNRM including a faking dimension are indeed associated with a more accurate representation of test-takers' substantive trait levels compared to only accounting for substantive traits and ERS in high-stakes assessment data (see LaHuis et al., 2019; Sun et al., 2022). Thus, modeling faking contributes to a purer assessment of interindividual differences. From an applied perspective, this can enhance test fairness and can help that decisions in high-stakes contexts, such as hiring decisions in personnel selection, are based on substantive trait score estimates that more closely reflect the traits of interest. Also, we found that latent correlations between substantive traits are debiased and more precisely estimated when faking is accounted for in the model. In applied settings, this implies that more valid conclusions on relationships between the assessed traits can be drawn and that, given that correlations between generally desirable traits are usually inflated in high-stakes assessments, more nuanced test-taker profiles within a personality inventory are possible.

Furthermore, the simulation showed that the extent to which a model including a faking dimension is superior to a model not including a faking dimension strongly depends on the impact of faking in the data. Unsurprisingly, a model with faking dimension outperformed a model without faking dimension more strongly when the faking impact in the data-generating process was high than when it was low. Crucially, however, a model with faking dimension was never inferior to a model without faking dimension and did not erroneously attribute substantive trait variance to a faking dimension, even when faking was completely absent in the data. Thus, using the model with faking dimension can be recommended in applied contexts in which faking might occur.

Importance of item desirability characteristics

Along with studying the mere psychometric benefits of the faking modeling approach, we examined the effects of different desirability characteristics of items. That is, we investigated how variation in the way item content is related to social desirability can facilitate the modeling of faking and counteract its detrimental effects. As Peabody (1967) noted, most personality items confound descriptive aspects with evaluative aspects, which implies that high scores can be due to a high substantive trait level, a high faking tendency, or both, unless faking is statistically accounted for. Concerning the MNRM, however, the confound between descriptive and evaluative aspects causes high collinearity between the scoring weight vectors of the substantive trait and faking dimensions, which arguably makes it difficult to properly disentangle substantive traits and faking.

In the simulation, it turned out that item desirability characteristics, on the one hand, moderate the effect of modeling faking and, on the other hand, have a main effect regarding the recovery of substantive trait person parameters and latent correlations between substantive traits. In particular, a model accounting for faking was differentially superior to a model ignoring faking depending on the composition of desirability trajectories. Additionally, more variety in desirability trajectories across items was associated with a generally better parameter recovery.

When the faking impact was low, the difference in parameter recovery between a model ignoring faking and a model accounting for faking was most pronounced in item compositions that only contained increasing desirability trajectories. Hence, despite collinearity between scoring weight vectors of substantive traits and faking in these item compositions, modeling faking can be particularly worthwhile compared to ignoring faking. Nevertheless, parameter recovery in item compositions with mainly increasing desirability trajectories was generally worse than in item compositions that also contained decreasing desirability trajectories. Here, a model ignoring faking could also recover parameters well. For the case of a low faking impact, the effects of faking on item responses hence seem to cancel each other out across the items of a test with balanced desirability trajectories, such that even ignoring faking can yield satisfactory results.

However, when the faking impact was high, modeling faking was notably superior to not modeling faking in all item compositions. Again, more variety in desirability trajectories was associated with a generally better parameter recovery, but a model without faking dimension only profited considerably from this in item compositions that contained all desirability trajectory types. A model with faking dimension, in contrast, entailed good parameter recovery even if increasing desirability trajectories were complemented only with inverted-U-shaped and/or nonmonotonically decreasing desirability trajectories. Hence, for the case of a high faking impact, modeling faking is necessary to achieve good parameter recovery irrespective of item desirability characteristics. Conceptually, effects of faking on item responses can also be expected to cancel each other out across items with balanced desirability trajectories when the faking impact is high. Because the effects of faking on item responses do not need to be constant across items, however, it is unlikely that the effects average out entirely within a given item set. Considering that this imperfect averaging-out should be more pronounced in case of a high faking impact, a model that accounts for item-specific effects of faking is required irrespective of item desirability characteristics to recover parameters well in this case.

Having modified items from the widely-used personality test BFI-2, we also demonstrated that it is possible to create more variety in empirical desirability trajectories through item refinement and that this is associated with the same result patterns as in the simulation. Hence, deconfounding descriptive and evaluative aspects in items of a personality test is not only appealing from a theoretical and conceptual point of view but is also feasible empirically and has utility for applied assessments. Resembling the findings from the simulation study, item desirability characteristics in the empirical demonstration interacted with the effect of modeling faking and had a main effect regarding the extent to which the estimates from the HS condition recovered the LS condition estimates. Despite collinearity between scoring weight vectors of substantive traits and faking, the effect of modeling faking was most pronounced in item compositions that only consisted of items with increasing desirability trajectories. More variety in desirability trajectories, in turn, reduced differences between modeling faking and ignoring faking, but led to an estimation of parameters that generally better recovered the LS condition estimates, mirroring the simulation findings from conditions with a low faking impact. As discussed above, this stands to reason considering the experimental nature of the present faking manipulation.

Note, however, the differences between the item modification approach of this article and other approaches of item redesign.[10] Bäckström et al. (2009, 2023) and Wood et al. (2022, 2023), for example, followed an approach of neutralizing item evaluativeness. Like our approach, this approach aims to reword items regarding desirability but preserve the substantive item content. However, whereas our approach seeks to create variation in desirability trajectories across items, the approach of item evaluativeness neutralization implies that either all response categories have the same (intermediate) level of desirability or that the midpoint of the rating scale has the highest desirability level. Items modified according to this approach are, however, not suited for modeling faking by means of the MNRM because they either have constant faking scoring weights for all categories or have scoring weight vectors of faking that are redundant to scoring weight vectors of other response biases. The former would make the faking dimension irrelevant, the latter would make it impossible to separate faking from response styles like midscale response style (MRS), which is the tendency to prefer the midpoint category of a rating scale irrespective of item content.

One might argue that neutralizing item evaluativeness can eliminate SDR and faking in the first place. However, it is questionable whether test-takers would indeed only respond according to their substantive traits once they are confronted with an evaluatively neutral item. Apart from the fact that items with reduced evaluativeness can still give test-takers information regarding desirability (Wood et al., 2023), it is likely that, even for perfectly neutral items, test-takers would still try to figure out what is desirable in the given assessment context and then edit responses according to their idiosyncratic conclusions. In turn, by modeling items with pretested desirability trajectories, the MNRM explicitly takes item-specific response editing into account and thus affords a model-based separation of substantive traits and faking. Additionally, it yields estimates of each test-taker's degree of faking in a given assessment context. Having such an estimate can be a helpful piece of information to evaluate the trustworthiness of responses from a test-taker in an applied assessment and can be used to study the substantive nature of the faking construct (Seitz et al., 2023).

[10] Peabody (Peabody, 1967, 1984) and other scholars (e.g., Borkenau & Ostendorf, 1989; Petterson et al., 2012; Saucier, 1994) proposed an approach of deconfounding descriptive and evaluative aspects that is similar to our approach. Namely, they created sets of four adjectives that are balanced with respect to the direction of descriptive and the valence of evaluative item content. Their studies, however, were limited to trait inferences from adjectives and did not feature typical personality test items in the form of statements that test-takers can more or less agree with. Also, they did not account for faking in a model-based manner using these kinds of items.

Limitations

Along with the above-described advantages of modeling faking by means of the MNRM in combination with different item desirability characteristics, some limitations should be mentioned. First, the fact that faking scoring weights for a specific item set and assessment situation are not readily transferable to other items or another assessment context can be considered a pragmatic limitation because additional resources will be required to adequately specify faking scoring weights if one wants to model different items or responses from different assessment settings. At the same time, the specificity of faking scoring weights can also be regarded as an asset of the present modeling approach because the modeling of faking is thus tailored to the specific assessment situation at hand.

Second, because scoring weights of faking are person-invariant model parameters, it is implicitly assumed that one desirability trajectory per item is appropriate to capture faking for all test-takers. However, as can be seen in the variability of desirability ratings in the pilot study, people do not perfectly agree about the most desirable category of an item. The more strongly test-takers differ in how they perceive desirability and respond according to it, the less appropriate it will be to use scoring weights of faking that are fixed across persons. To incorporate individual desirability perceptions of test-takers, however, one would have to collect additional data from the same test-takers whose actual item responses are to be modeled, which has multiple methodological shortcomings and will often not be possible in practice. Instead, the model makes the assumption that individual deviations in desirability perceptions from the specified item desirability characteristics are unsystematic fluctuations around a desirability trajectory that is on average representative for all test-takers. Thus, the model uses average desirability perceptions concerning each item to account for faking along with substantive traits and other response biases. This extends previous faking modeling approaches that assume effect patterns of faking to be constant across both persons and items (e.g., Böckenholt, 2014; Brown & Böckenholt, 2022; Hendy et al., 2021; Leng et al., 2020; Ziegler & Bühner, 2009). Nevertheless, to keep systematic deviations between the specified desirability characteristics and test-takers' real desirability perceptions minimal, we advise researchers and practitioners to collect desirability ratings from a pilot study sample that is maximally similar to the sample of actual test-takers in terms of demographic features and contextual factors. Also, future research should investigate how much disparity in test-takers' desirability perceptions is acceptable for the presented faking modeling approach to produce satisfactory results. For such robustness checks of the model, further simulation studies would be appropriate to determine a criterion for the necessary level of agreement in individual desirability perceptions. At the same time, if one knows about systematic deviations of desirability perceptions between groups of test-takers (e.g., young professionals vs. experienced hires), future studies could also specify scoring weights of faking group-specifically.

Third, modifying items to create more variety in desirability trajectories can be a challenging endeavor for items of substantive traits that are inherently desirable or undesirable in a given context. As Wood et al. (2022) noted, social desirability of personality items can be "partially intrinsic and partially the result of item writing practices" (p. 818). That is, for personality traits that are intrinsically desirable (undesirable), it can be hard to generate items with inverted-U-shaped and/or decreasing (increasing) desirability trajectories and, at the same time, not change the meaning of the assessed constructs. The risk of subtly changing the meaning of the construct applies to all kinds of item rewording, but especially to the attempt of creating items with decreasing (increasing) desirability trajectories for personality traits that intrinsically intertwine substantive and desirable (undesirable) attributes. To meet this problem, it is vital to review modified items regarding their fit to the construct definitions and to only include items in the final test form that still discriminate well concerning the traits of interest. In the empirical demonstration of this article, we did so by selecting the items with highest discrimination concerning the underlying Big Five trait in the condition in which participants were instructed to respond honestly. Nevertheless, as can be seen in Table S.II.1, convergent validities with the BFI-2 dropped to some extent in item compositions that also contained decreasing desirability trajectories compared to item compositions that only comprised increasing desirability trajectories. Specifically, the correlations between the sum score from the original BFI-2 items and the sum scores from the five item compositions ranged from .94 to .75 for Extraversion, from .92 to .65 for Agreeableness, from .97 to .63 for Conscientiousness, from .95 to .72 for Emotional Stability, and from .93 to .84 for Openness. However, these values are still in the upper range of typical convergent validities between different Big Five tests that feature distinct facets and emphases. Danner et al. (2019), for instance, reported correlations between the BFI-2 and other popular Big Five tests ranging from .88 to .64. Soto and John (2017) similarly found convergent validities ranging from .94 to .68. Thus, in the empirical demonstration of this article, changes in convergent validities associated with the modification of item desirability characteristics were empirically no larger than must be expected when switching from one Big Five test to another. Concerning applied settings, we argue that it mainly depends on the researcher's or practitioner's goals how much change in the meaning of the construct can be accepted. In personality research contexts, where the primary goal is to measure personality traits that are narrowly defined by particular facets, modifying item desirability characteristics might be less appropriate than in applied measurement contexts, such as high-stakes assessments in personnel selection, where the primary goal is to have a fair assessment that is not contaminated by faking. The more important this latter goal is, the more will subtle changes in the construct meaning be offset by having a measure that is not easily fakable, especially if the scoring of the test is based on the presented IRT model where different item desirability characteristics are explicitly modeled by the faking dimension.

Future research directions

In the faking modeling approach presented in this article, faking is conceptualized as a continuous individual difference variable. Even though treating faking as such an individual difference variable is consistent with Ziegler et al.'s (2015) finding that faking mainly represents a continuous variable as opposed to a manifestation of distinct response processes, there might be heterogeneity in response strategies over and above quantitative variation in the degree of faking. To further examine the nature of heterogeneity in faking, future research could extend the model of this article in a mixture modeling framework by allowing for latent classes characterized by qualitatively different response processes. Also, faking might be better described by distributions other than a normal distribution. Future studies could, for instance, model faking using a truncated normal or log-normal distribution. This would correspond to a conceptualization of faking as a unipolar construct. Recent IRT approaches for unipolar modeling of performance data or psychopathological constructs (e.g., Huang & Bolt, 2023; Lucke, 2015) could be used as a starting point for future model extensions in this regard, though such models are currently limited to the case of modeling a single latent dimension.

Follow-up studies could also examine how the approach of changing item desirability characteristics affects the prediction of outcomes that are of interest in high-stakes assessments, such as job performance. Different effects are conceivable: First, considering that studies have often found a limited influence of SDR and faking on predictive validity (e.g., Ones et al., 2007; Paunonen & LeBel, 2012), it could be that correlations with outcomes are not affected. Second, given that part of the desirable item content can be beneficial for predicting performance outcomes (e.g., Li & Bagger, 2006; Wood et al., 2023), it could be that the prediction of these outcomes deteriorates because the modified item sets also capture less desirable aspects of the traits of interest. Third, it could be that the prediction improves, assuming that faking acts as a suppressor in predicting outcomes (e.g., Bing et al., 2011; Hakstian & Ng, 2005) and that substantive trait scores are less distorted by faking once there is more variety in desirability trajectories.

Moreover, since the current study featured an experimental faking manipulation, future studies should replicate the results in high-stakes assessment data from the field, for instance, in a dataset that contains responses from the same test-takers as job applicants and as job incumbents. However, compared to other studies in which faking was induced experimentally, the present study aimed at approximating the circumstances of an actual application context in two ways: First, the faking manipulation in this study instructed participants to base responses on their actual attitudes and behaviors (which is a typical response instruction in a personnel selection context) but at the same time try to get the vacant position (which is the goal of people applying for a particular job). In the faking literature, however, it is not uncommon to find blatant faking instructions in which participants are simply told to respond in a socially desirable manner. Second, there was a financial incentive for response distortion that created actual stakes for participants. Assessment results thus carried real consequences, emulating the incentive structure of a high-stakes testing like in personnel selection (i.e., a dilemma between sticking to the instruction of responding honestly and giving distorted responses to receive the reward). Nonetheless, follow-up studies with data from actual personnel selection contexts would be welcome. Thereby, it would also be appealing to further validate the model's adjustments of substantive trait scores with personality measures that are less susceptible to faking, such as multidimensional forced-choice (MFC) measures (e.g., Cao & Drasgow, 2019) or observer ratings of personality (e.g., König et al., 2017).

Conclusion

To conclude, the present research demonstrates two interacting approaches to address the response bias of faking: First, the MNRM provides an appealing framework for statistically modeling the influence of faking on item responses, which is particularly effective when the faking impact in the data is high. Second, modifying desirability characteristics of items can be a means to facilitate the modeling of faking and to counteract its adverse effects in the first place. Furthermore, this article highlights circumstances under which a statistical modeling of faking is particularly important and useful to improve the assessment of psychological constructs, and it reveals the beneficial effects of considering item desirability characteristics already at the stage of item construction to remedy the negative psychometric effects of faking. Our findings provide guidelines for applied researchers and practitioners to decide when using the MNRM to model faking is worthwhile and how to address faking by refining self-report personality questionnaires.

Appendix: List of items used in the empirical demonstration

Table A1 BFI-2 and modified items with their respective desirability trajectory classification

Item code	German version	English translation	Desirability trajectory classification
BFI-2 items, Extraversion:			
BFI_E01	Ich gehe aus mir heraus, bin gesellig.	I am outgoing, sociable.	nonmonotonically increasing
BFI_E02	Ich bin eher schüchtern. (R)	I am rather shy. (R)	monotonically increasing
BFI_E03	Ich bin eher ruhig. (R)	I am rather quiet. (R)	inverted-U-shaped
BFI_E04	Ich bin gesprächig.	I am talkative.	nonmonotonically increasing
BFI_E05	Ich bin durchsetzungsfähig, energisch.	I am assertive, energetic.	monotonically increasing
BFI_E06	Ich neige dazu, die Führung zu übernehmen.	I tend to act as a leader.	monotonically increasing
BFI_E07	Mir fällt es schwer, andere zu beeinflussen. (R)	I find it hard to influence people. (R)	nonmonotonically increasing
BFI_E08	In einer Gruppe überlasse ich lieber anderen die Entscheidung. (R)	In a group, I prefer to have others take charge. (R)	monotonically increasing
BFI_E09	Ich schäume selten vor Begeisterung über (R).	I rarely feel excited or eager. (R)	nonmonotonically increasing
BFI_E10	Ich bin weniger aktiv und unternehmungslustig als andere. (R)	I am less active and adventurous than other people. (R)	nonmonotonically increasing
BFI_E11	Ich bin voller Energie und Tatendrang.	I am full of energy and drive.	monotonically increasing
BFI_E12	Ich bin begeisterungsfähig und kann andere leicht mitreißen.	I am enthusiastic and can easily carry others along.	monotonically increasing
Modified items, Extraversion:			
mod_E01	Ich brauche ständigen Kontakt zu anderen Menschen.	I need constant contact with other people.	inverted-U-shaped
mod_E02	Mir fällt es leicht, auch einmal zu schweigen. (R)	It is easy for me to remain silent once in a while. (R)	nonmonotonically decreasing
mod_E03	Ich verwickle andere gerne in sehr lange Gespräche.	I like to engage others in very long conversations.	nonmonotonically decreasing
mod_E04	Ich bin geschwätzig.	I am chatty.	nonmonotonically decreasing
mod_E05	Ich bin so redselig, dass ich anderen damit manchmal auf die Nerven gehe.	I am so talkative that sometimes I annoy other people.	monotonically decreasing
mod_E06	Ich stehe ungern im Mittelpunkt des Interesses. (R)	I don't like to be the center of interest. (R)	nonmonotonically increasing
mod_E07	Ich ziehe gerne die Aufmerksamkeit auf mich.	I like to draw attention to myself.	inverted-U-shaped
mod_E08	Bei Gruppenprojekten stehe ich meistens nicht im Mittelpunkt. (R)	I am usually not the center of attention in group projects. (R)	nonmonotonically increasing
mod_E09	Für gewöhnlich dominiere ich Gespräche.	I usually dominate conversations.	inverted-U-shaped
mod_E10	Ich kann Freude daran haben, nicht aktiv zu sein. (R)	I can find joy in not being active. (R)	nonmonotonically increasing
mod_E11	In einer Gruppe bin ich bei jeder Aktivität dabei.	In a group, I participate in every activity.	inverted-U-shaped
mod_E12	Mein Tatendrang überfordert andere manchmal.	My drive for action sometimes overwhelms others.	inverted-U-shaped
mod_E13	Mit meiner Begeisterung schieße ich gelegentlich über das Ziel hinaus.	I occasionally overshoot the mark with my enthusiasm.	inverted-U-shaped
BFI-2 items, Agreeableness:			
BFI_A01	Ich bin einfühlsam, warmherzig.	I am compassionate, warm-hearted.	nonmonotonically increasing
BFI_A02	Ich habe mit anderen wenig Mitgefühl. (R)	I have little sympathy for others. (R)	monotonically increasing
BFI_A03	Ich bin hilfsbereit und selbstlos.	I am helpful and unselfish with others.	nonmonotonically increasing
BFI_A04	Andere sind mir eher gleichgültig, egal. (R)	Others are of no concern and inconsequential to me. (R)	monotonically increasing
BFI_A05	Ich begegne anderen mit Respekt.	I treat others with respect.	monotonically increasing
BFI_A06	Ich habe oft Streit mit anderen. (R)	I often have arguments with others. (R)	monotonically increasing
BFI_A07	Ich bin manchmal unhöflich und schroff. (R)	I am sometimes rude and harsh. (R)	monotonically increasing

Table A1 (continued)

Item code	German version	English translation	Desirability trajectory classification
BFI_A08	Ich bin höflich und zuvorkommend.	I am polite and courteous.	monotonically increasing
BFI_A09	Ich neige dazu, andere zu kritisieren. (R)	I tend to criticize others. (R)	nonmonotonically increasing
BFI_A10	Ich bin nachsichtig, vergebe anderen leicht.	I am indulgent and have a forgiving nature.	inverted-U-shaped
BFI_A11	Ich bin anderen gegenüber misstrauisch. (R)	I am suspicious of others. (R)	nonmonotonically increasing
BFI_A12	Ich schenke anderen leicht Vertrauen, glaube an das Gute im Menschen.	I trust others easily and assume the best about people.	inverted-U-shaped
Modified items, Agreeableness:			
mod_A01	Ich leide mit den Problemen anderer sehr stark mit.	I am very strongly affected by other people's problems.	nonmonotonically decreasing
mod_A02	Ich verbringe viel Zeit damit, mich um die Bedürfnisse anderer zu kümmern.	I spend a lot of time taking care of other people's needs.	inverted-U-shaped
mod_A03	Ich kann mich gut von den Emotionen anderer distanzieren. (R)	I am good at distancing myself from the emotions of others. (R)	nonmonotonically decreasing
mod_A04	Durch die Probleme anderer lasse ich mich nicht von meinen Zielen abbringen. (R)	I don't let the problems of others distract me from my goals. (R)	nonmonotonically decreasing
mod_A05	Ich kann nur schwer Entscheidungen treffen, welche andere verletzen könnten.	I find it difficult to make decisions that could hurt others.	nonmonotonically decreasing
mod_A06	Ich kann anderen Personen nur schwer einen Wunsch ausschlagen.	I find it difficult to refuse other people a wish.	nonmonotonically decreasing
mod_A07	Unangenehme Gespräche zu führen, macht mir nichts aus. (R)	I don't mind having uncomfortable conversations. (R)	monotonically decreasing
mod_A08	Ich scheue mich nicht vor hitzigen Diskussionen. (R)	I don't shy away from heated discussions. (R)	nonmonotonically decreasing
mod_A09	Ich gebe oft nach, um Streit zu vermeiden.	I often give in to avoid arguments.	nonmonotonically decreasing
mod_A10	Ich gebe lieber nach, als eine Meinungsverschiedenheit auszudiskutieren.	I would rather give in than argue out a difference of opinion.	monotonically decreasing
mod_A11	Ich bin ein sehr harmoniebedürftiger Mensch.	I am a very harmony-seeking person.	inverted-U-shaped
mod_A12	Kritik an anderen zu äußern, fällt mir nicht schwer. (R)	I don't find it difficult to criticize others. (R)	nonmonotonically decreasing
mod_A13	Ich bin sehr nachsichtig.	I am very indulgent.	inverted-U-shaped
mod_A14	Mir fällt es schwer, auch einmal „nein" zu sagen.	I find it hard to say "no" once in a while.	monotonically decreasing
BFI-2 items, Conscientiousness:			
BFI_C01	Ich bin eher unordentlich. (R)	I am rather messy. (R)	monotonically increasing
BFI_C02	Ich bin systematisch, halte meine Sachen in Ordnung.	I am systematic and keep things in order.	nonmonotonically increasing
BFI_C03	Ich mag es sauber und aufgeräumt.	I like to keep things neat and tidy.	nonmonotonically increasing
BFI_C04	Ich bin eher der chaotische Typ, mache selten sauber. (R)	I am more of a chaotic type and rarely clean up. (R)	monotonically increasing
BFI_C05	Ich bin bequem, neige zu Faulheit. (R)	I tend to be lazy. (R)	monotonically increasing
BFI_C06	Ich neige dazu, Aufgaben vor mir herzuschieben. (R)	I have difficulty getting started on tasks. (R)	monotonically increasing
BFI_C07	Ich bin effizient, erledige Dinge schnell.	I am efficient and get things done quickly.	monotonically increasing
BFI_C08	Ich bleibe an einer Aufgabe dran, bis sie erledigt ist.	I am persistent and work until the task is finished.	monotonically increasing
BFI_C09	Ich bin stetig, beständig.	I am dependable, steady.	nonmonotonically increasing
BFI_C10	Ich bin manchmal ziemlich nachlässig. (R)	I can occasionally be somewhat careless. (R)	monotonically increasing
BFI_C11	Ich bin verlässlich, auf mich kann man zählen.	I am a reliable person one can count on.	monotonically increasing
BFI_C12	Manchmal verhalte ich mich verantwortungslos, leichtsinnig. (R)	I sometimes behave irresponsibly and recklessly. (R)	monotonically increasing
Modified items, Conscientiousness:			
mod_C01	Ich verliere viel Zeit damit, meine Sachen zu ordnen.	I lose a lot of time organizing my things.	monotonically decreasing

Table A1 (continued)

Item code	German version	English translation	Desirability trajectory classification
mod_C02	Beim Thema Ordnung bin ich nicht pingelig. (R)	I am not picky when it comes to tidiness. (R)	nonmonotonically increasing
mod_C03	Ich bin penibel.	I am fastidious.	inverted-U-shaped
mod_C04	Ich bin als Perfektionist bekannt.	I am known as a perfectionist.	inverted-U-shaped
mod_C05	Dem Ordnen von Dokumenten und Dateien räume ich viel Zeit ein.	I spend a lot of time organizing documents and files.	inverted-U-shaped
mod_C06	Von Plänen weiche ich nur ungern ab.	I don't like to deviate from plans.	inverted-U-shaped
mod_C07	Wenn ich Aufgaben nicht sofort erledigen kann, fühle ich mich schlecht.	If I can't complete tasks immediately, I feel bad.	inverted-U-shaped
mod_C08	Ich kann unabgeschlossene Projekte auch einmal für einige Zeit ruhen lassen. (R)	I can also let unfinished projects rest for a while. (R)	inverted-U-shaped
mod_C09	Ich verbeiße mich in Aufgaben, bis ich zu einer Lösung gelange.	I get wound up in tasks until I reach a solution.	nonmonotonically increasing
mod_C10	Alltägliche Aufgaben erledige ich so sorgfältig, dass ich oft länger brauche als erforderlich.	I perform everyday tasks so thoroughly that I often need longer than necessary.	nonmonotonically decreasing
mod_C11	Selbst wenn ich einen belanglosen Fehler mache, kann ich diesen nur schwer akzeptieren.	Even if I make a trivial mistake, I find it hard to accept.	monotonically decreasing
mod_C12	Auch bei unwichtigen Projekten arbeite ich sehr akribisch.	I work very meticulously even on unimportant projects.	nonmonotonically increasing
mod_C13	Es macht mir nichts aus, Dinge aufzuschieben. (R)	I don't mind putting things off. (R)	nonmonotonically increasing
BFI-2 items, Emotional Stability:			
BFI_N01	Ich bleibe auch in stressigen Situationen gelassen. (R)	I stay calm even in stressful situations. (R)	monotonically increasing
BFI_N02	Ich reagiere leicht angespannt.	I easily react tensely.	monotonically increasing
BFI_N03	Ich mache mir oft Sorgen.	I worry a lot.	nonmonotonically increasing
BFI_N04	Ich werde selten nervös und unsicher. (R)	I rarely feel anxious and insecure. (R)	monotonically increasing
BFI_N05	Ich bleibe auch bei Rückschlägen zuversichtlich. (R)	I stay confident after experiencing a setback. (R)	monotonically increasing
BFI_N06	Ich bin selbstsicher, mit mir zufrieden. (R)	I am self-confident and content with me. (R)	monotonically increasing
BFI_N07	Ich fühle mich oft bedrückt, freudlos.	I often feel sad, joyless.	monotonically increasing
BFI_N08	Ich bin oft deprimiert, niedergeschlagen.	I tend to feel depressed, blue.	monotonically increasing
BFI_N09	Ich kann launisch sein, habe schwankende Stimmungen.	I can be moody and have up-and-down mood swings.	monotonically increasing
BFI_N10	Ich bin ausgeglichen, nicht leicht aus der Ruhe zu bringen. (R)	I am even-tempered, not easily upset. (R)	monotonically increasing
BFI_N11	Ich habe meine Gefühle unter Kontrolle, werde selten wütend. (R)	I have my emotions under control and rarely get angry. (R)	monotonically increasing
BFI_N12	Ich reagiere schnell gereizt oder genervt.	I quickly become irritated or annoyed.	monotonically increasing
Modified items, Emotional Stability:			
mod_N01	Kaum etwas kann meinen emotionalen Zustand verändern. (R)	Hardly anything can change my emotional state. (R)	inverted-U-shaped
mod_N02	Meine Stimmung hängt nicht von äußeren Umständen ab. (R)	My mood does not depend on external circumstances. (R)	monotonically increasing
mod_N03	Auf bedeutsame Ereignisse reagiere ich unemotional. (R)	I react unemotionally to significant events. (R)	nonmonotonically decreasing
mod_N04	Ich bin sensibel für meine Gefühle und Stimmungen.	I am sensitive to my feelings and moods.	inverted-U-shaped
mod_N05	Selbst in gefährlichen Situationen verspüre ich keine Angst. (R)	Even in dangerous situations, I feel no fear. (R)	inverted-U-shaped
mod_N06	Ich erkenne Risiken sehr früh.	I recognize risks very early.	monotonically decreasing

Table A1 (continued)

Item code	German version	English translation	Desirability trajectory classification
mod_N07	Nichts kann dazu führen, dass ich niedergeschlagen bin. (R)	Nothing can cause me to be dejected. (R)	inverted-U-shaped
mod_N08	Meinungsverschiedenheiten können mich nach Feierabend weiter verfolgen.	Differences of opinion can continue to haunt me after work.	nonmonotonically increasing
mod_N09	Kritik an meiner Arbeit lässt mich kalt. (R)	Criticism of my work leaves me cold. (R)	nonmonotonically decreasing
mod_N10	Persönliche Kritik kann mir nichts anhaben. (R)	Personal criticism cannot harm me. (R)	nonmonotonically increasing
mod_N11	Ich empfinde selten starke Emotionen. (R)	I rarely feel strong emotions. (R)	inverted-U-shaped
mod_N12	Auf der Arbeit ist es noch niemandem gelungen, mich emotional zu verletzen. (R)	No one at work has ever managed to hurt me emotionally. (R)	nonmonotonically increasing
mod_N13	Gelegentlich merke ich, dass ich leicht verletzlich bin.	Occasionally I notice that I am slightly vulnerable.	nonmonotonically increasing
BFI-2 items, Openness:			
BFI_O01	Ich bin nicht sonderlich kunstinteressiert. (R)	I am not particularly interested in art. (R)	inverted-U-shaped
BFI_O02	Ich kann mich für Kunst, Musik und Literatur begeistern.	I can be fascinated by art, music, and literature.	inverted-U-shaped
BFI_O03	Ich weiß Kunst und Schönheit zu schätzen.	I value art and beauty.	inverted-U-shaped
BFI_O04	Ich finde Gedichte und Theaterstücke langweilig. (R)	I think poetry and plays are boring. (R)	inverted-U-shaped
BFI_O05	Ich bin vielseitig interessiert.	I have a wide range of interests.	monotonically increasing
BFI_O06	Ich meide philosophische Diskussionen. (R)	I avoid philosophical discussions. (R)	inverted-U-shaped
BFI_O07	Es macht mir Spaß, gründlich über komplexe Dinge nachzudenken und sie zu verstehen.	I enjoy thinking deeply about complex things and understanding them.	monotonically increasing
BFI_O08	Mich interessieren abstrakte Überlegungen wenig. (R)	I have little interest in abstract ideas. (R)	nonmonotonically increasing
BFI_O09	Ich bin erfinderisch, mir fallen raffinierte Lösungen ein.	I am inventive and find clever ways to do things.	monotonically increasing
BFI_O10	Ich bin nicht besonders einfallsreich. (R)	I have little creativity. (R)	monotonically increasing
BFI_O11	Ich bin nicht sonderlich fantasievoll. (R)	I have difficulty imagining things. (R)	nonmonotonically increasing
BFI_O12	Ich bin originell, entwickle neue Ideen.	I am original and come up with new ideas.	monotonically increasing
Modified items, Openness:			
mod_O01	Ich kann mich in Kunst, Musik und Literatur verlieren.	I can lose myself in art, music, and literature.	nonmonotonically decreasing
mod_O02	Ich kann Fantasien nicht viel abgewinnen. (R)	I don't take much pleasure in fantasizing. (R)	nonmonotonically increasing
mod_O03	Ich setze Projekte lieber praktisch um, als mich mit theoretischen Aspekten zu beschäftigen. (R)	I prefer to put projects into practice than deal with theoretical aspects. (R)	inverted-U-shaped
mod_O04	Routinetätigkeiten langweilen mich schnell.	Routine tasks bore me quickly.	nonmonotonically decreasing
mod_O05	Neue Aufgaben ziehe ich Tätigkeiten vor, mit denen ich mich auskenne.	I prefer new tasks to activities that I am familiar with.	inverted-U-shaped
mod_O06	Bei neuen Problemen greife ich auf altbewährte Methoden zurück. (R)	I fall back on tried and tested methods when faced with new problems. (R)	inverted-U-shaped
mod_O07	Ich finde jedes Mal einen neuen Weg, an eine bekannte Aufgabe heranzugehen.	I always find a new approach to a familiar task.	nonmonotonically increasing
mod_O08	Ich verweile nicht lange in Träumen und Fantasien. (R)	I don't linger in dreams and fantasies for long. (R)	nonmonotonically decreasing
mod_O09	Ich habe oft träumerische Gedanken.	I often have dreamy thoughts.	nonmonotonically decreasing
mod_O10	Meine Herangehensweisen an Aufgaben sind oft unkonventionell.	My approach to tasks is often unconventional.	inverted-U-shaped
mod_O11	Ich bin sehr experimentierfreudig.	I am very keen to experiment.	nonmonotonically increasing
mod_O12	Meine Ideen sind oftmals weit hergeholt.	My ideas are often far-fetched.	nonmonotonically decreasing

Note. The classification of desirability trajectories for negatively-keyed items is for the case of recoded item responses. BFI-2 = *Big Five Inventory 2* (Danner et al., 2016, 2019); (R) = negatively-keyed items

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-024-02509-x.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2277 "Statistical Modeling in Psychology" (SMiP).

Availability of data and materials All data and materials are available at https://osf.io/ms57p/.

Availability of code R scripts for the simulation and the empirical analyses as well as detailed explanations for specifying and estimating the MNRM are available at https://osf.io/ms57p/.

Declarations

Conflicts of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethics approval No ethics approval was necessary according to German laws because the study did not involve the deception of participants.

Consent to participate Informed consent was obtained from all participants of the study.

Consent for publication Only participants who did not indicate that their anonymized data shall not be used for analysis and publication were included.

Preregistration Analyses in this research were not preregistered because the purpose of the study was the evaluation of a psychometric model through a statistical simulation. The part of the study involving empirical data served as an illustration of the simulation findings in empirical questionnaire data and did not involve a confirmatory testing of hypotheses.

References

Ames, D. R., & Flynn, F. J. (2007). What breaks a leader: The curvilinear relation between assertiveness and leadership. *Journal of Personality and Social Psychology, 92*(2), 307–324. https://doi.org/10.1037/0022-3514.92.2.307

Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*(3), 335–344. https://doi.org/10.1016/j.jrp.2008.12.013

Bäckström, M., Björklund, F., Maddux, R. E., & Lindén, M. (2023). The NB5I: A full-scale Big-Five inventory with evaluatively neutralized items. *European Journal of Psychological Assessment, 39*(2), 132–140. https://doi.org/10.1027/1015-5759/a000687

Baron, H., Gibbons, P., MacIver, R., & Nyfield, G. (2000). Conscientiousness and managerial performance. *Journal of Occupational and Organizational Psychology, 73*(2), 171–180. https://doi.org/10.1348/096317900166967

Bing, M. N., Kluemper, D., Kristl Davison, H., Taylor, S., & Novicevic, M. (2011). Overclaiming as a measure of faking. *Organizational Behavior and Human Decision Processes, 116*(1), 148–162. https://doi.org/10.1016/j.obhdp.2011.05.006

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. https://doi.org/10.1007/bf02291411

Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. *Psychometrika, 79*(3), 515–537. https://doi.org/10.1007/s11336-013-9390-9

Bolt, D. M., & Newton, J. R. (2011). Multiscale measurement of extreme response style. *Educational and Psychological Measurement, 71*(5), 814–833. https://doi.org/10.1177/0013164410388411

Borkenau, P., & Ostendorf, F. (1989). Descriptive consistency and social desirability in self- and peer reports. *European Journal of Personality, 3*(1), 31–45. https://doi.org/10.1002/per.2410030105

Brown, A., & Böckenholt, U. (2022). Intermittent faking of personality profiles in high-stakes assessments: A grade of membership analysis. *Psychological Methods, 27*(5), 895–916. https://doi.org/10.1037/met0000295

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics, 35*(3), 307–335. https://doi.org/10.3102/1076998609353115

Cai, L., & Monroe, S. (2013). IRT model fit evaluation from theory to practice: Progress and some unanswered questions. *Measurement: Interdisciplinary Research & Perspective, 11*(3), 102–106. https://doi.org/10.1080/15366367.2013.835172

Cai, L., & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data* (CRESST Report 839). National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California.

Cao, M., & Drasgow, F. (2019). Does forcing reduce faking? A meta-analytic review of forced-choice personality measures in high-stakes situations. *Journal of Applied Psychology, 104*(11), 1347–1368. https://doi.org/10.1037/apl0000414

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

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology, 16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C., & John, O. P. (2016). Die deutsche Version des Big Five Inventory 2 (BFI-2) [The German version of the Big Five Inventory 2 (BFI-2)]. *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)* [*Compilation of items and scales for the social sciences (ZIS)*]. https://doi.org/10.6102/zis247

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C. J., & John, O. P. (2019). Das Big Five Inventar

2: Validierung eines Persönlichkeitsinventars zur Erfassung von 5 Persönlichkeitsdomänen und 15 Facetten [The Big Five Inventory 2: Validation of a personality inventory for measuring 5 personality domains 15 facets]. *Diagnostica, 65*(3), 121–132. https://doi.org/10.1026/0012-1924/a000218

De Hoogh, A. H. B., Den Hartog, D. N., & Koopman, P. L. (2005). Linking the Big Five-factors of personality to charismatic and transactional leadership; perceived dynamic work environment as a moderator. *Journal of Organizational Behavior, 26*(7), 839–865. https://doi.org/10.1002/job.344

de Vries, R. E., Zettler, I., & Hilbig, B. E. (2014). Rethinking trait conceptions of social desirability scales: Impression management as an expression of honesty-humility. *Assessment, 21*(3), 286–299. https://doi.org/10.1177/1073191113504619

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance, 16*(1), 81–106. https://doi.org/10.1207/S15327043HUP1601_4

Ellingson, J. E., & McFarland, L. A. (2011). Understanding faking behavior through the lens of motivation: An application of VIE theory. *Human Performance, 24*(4), 322–337. https://doi.org/10.1080/08959285.2011.597477

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*(2), 155–166. https://doi.org/10.1037/0021-9010.84.2.155

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

Falk, C. F., & Ju, U. (2020). Estimation of response styles using the multidimensional nominal response model: A tutorial and comparison with sum scores. *Frontiers in Psychology, 11*, 72. https://doi.org/10.3389/fpsyg.2020.00072

Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment, 11*(4), 340–344. https://doi.org/10.1111/j.0965-075X.2003.00256.x

Goldammer, P., Stöckli, P. L., Escher, Y. A., Annen, H., & Jonas, K. (2023). On the utility of indirect methods for detecting faking. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644231209520 Advance online publication

Griffith, R. L., & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0018

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*(3), 341–355. https://doi.org/10.1108/00483480710731310

Gummer, T., Roßmann, J., & Silber, H. (2021). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research, 50*(1), 238–264. https://doi.org/10.1177/0049124118769083

Hakstian, A. R., & Ng, E.-L. (2005). Employment-related motivational distortion: Its nature, measurement, and reduction. *Educational and Psychological Measurement, 65*(3), 405–441. https://doi.org/10.1177/0013164404267293

Hendy, N., Krammer, G., Schermer, J. A., & Biderman, M. D. (2021). Using bifactor models to identify faking on Big Five questionnaires. *International Journal of Selection and Assessment, 29*(1), 81–99. https://doi.org/10.1111/ijsa.12316

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

Hogan, R., & Kaiser, R. B. (2005). What we know about leadership. *Review of General Psychology, 9*(2), 169–180. https://doi.org/10.1037/1089-2680.9.2.169

Huang, Q., & Bolt, D. M. (2023). Unipolar IRT and the Author Recognition Test (ART). *Behavior Research Methods*.https://doi.org/10.3758/s13428-023-02275-2. Advance online publication.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*(4), 765–780. https://doi.org/10.1037/0021-9010.87.4.765

Kaiser, R. B., LeBreton, J. M., & Hogan, J. (2015). The dark side of personality and extreme leader behavior. *Applied Psychology, 64*(1), 55–92. https://doi.org/10.1111/apps.12024

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25*(4), 273–302. https://doi.org/10.1080/08959285.2012.703733

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology, 93*(1), 140–154. https://doi.org/10.1037/0021-9010.93.1.140

König, C. J., Hafsteinsson, L. G., Jansen, A., & Stadelmann, E. H. (2011). Applicants' self-presentational behavior across cultures: Less self-presentation in Switzerland and Iceland than in the United States. *International Journal of Selection and Assessment, 19*(4), 331–339. https://doi.org/10.1111/j.1468-2389.2011.00562.x

König, C. J., Steiner Thommen, L. A., Wittwer, A., & Kleinmann, M. (2017). Are observer ratings of applicants' personality also faked? Yes, but less than self-reports. *International Journal of Selection and Assessment, 25*(2), 183–192. https://doi.org/10.1111/ijsa.12171

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x

LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods, 12*(2), 296–319. https://doi.org/10.1177/1094428107302903

LaHuis, D. M., Blackmore, C. E., Bryant-Lees, K. B., & Delgado, K. (2019). Applying item response trees to personality data in the selection context. *Organizational Research Methods, 22*(4), 1007–1018. https://doi.org/10.1177/1094428118780310

Leiner, D. J. (2019). Too Fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods, 13*(3), 229–248. https://doi.org/10.18148/SRM/2019.V13I3.7403

Leng, C. H., Huang, H. Y., & Yao, G. (2020). A social desirability item response theory model: Retrieve-deceive-transfer. *Psychometrika, 85*(1), 56–74. https://doi.org/10.1007/s11336-019-09689-y

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*(2), 131–141. https://doi.org/10.1111/j.1468-2389.2006.00339.x

Lucke, J. F. (2015). Unipolar item response models. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 290–302). Routledge.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/bf02296272

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328. https://doi.org/10.1080/00273171.2014.911075

McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology, 51*(6), 882–888. https://doi.org/10.1037/0022-006x.51.6.882

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81–90. https://doi.org/10.1037/0022-3514.52.1.81

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*(5), 812–821. https://doi.org/10.1037/0021-9010.85.5.812

Merhof, V., Böhm, C. M., & Meiser, T. (2023). Separation of traits and extreme response style in IRTree models: The role of mimicry effects for the meaningful interpretation of Estimates. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644231213319. Advance online publication.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Müller, S., & Moshagen, M. (2019). True virtue, self-presentation, or both?: A behavioral test of impression management and overclaiming. *Psychological Assessment, 31*(2), 181–191. https://doi.org/10.1037/pas0000657

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

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*(4), 995–1027. https://doi.org/10.1111/j.1744-6570.2007.00099.x

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology, 46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Erlbaum.

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*(1), 158–175. https://doi.org/10.1037/a0028165

Peabody, D. (1967). Trait inferences: Evaluative and descriptive aspects. *Journal of Personality and Social Psychology, 7*(4, Pt. 2), 1–18. https://doi.org/10.1037/h0025230

Peabody, D. (1984). Personality dimensions through trait inferences. *Journal of Personality and Social Psychology, 46*(2), 384–403. https://doi.org/10.1037/0022-3514.46.2.384

Pettersson, E., Turkheimer, E., Horn, E. E., & Menatti, A. R. (2012). The general factor of personality and evaluation. *European Journal of Personality, 26*(3), 292–302. https://doi.org/10.1002/per.839

Saucier, G. (1994). Separating description and evaluation in the structure of personality attributes. *Journal of Personality and Social Psychology, 66*(1), 141–154. https://doi.org/10.1037/0022-3514.66.1.141

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*(6), 966–974. https://doi.org/10.1037/0021-9010.78.6.966

Seitz, T., Spengler, M., & Meiser, T. (2023, September 29). "What if applicants fake their responses?": Modeling faking in high-stakes in personality assessments using the multidimensional nominal response model. *PsyArXiv*. https://doi.org/10.31234/osf.io/j5mze

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2023). *afex: Analysis of factorial experiments* (version 1.3-0) [Computer software]. https://cran.r-project.org/web/packages/afex/index.html

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Sun, T., Zhang, B., Cao, M., & Drasgow, F. (2022). Faking detection improved: Adopting a Likert item response process tree model. *Organizational Research Methods, 25*(3), 490–512. https://doi.org/10.1177/10944281211002904

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408. https://doi.org/10.1007/bf02294363

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume One: Models* (pp. 51–73). Chapman & Hall/CRC Press.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577. https://doi.org/10.1007/BF02295596

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Lawrence Erlbaum Associates Publishers. https://www.taylorfrancis.com/books/edit/10.4324/9781410604729/test-scoring-david-thissen-howard-wainer

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality, 44*(3), 315–327. https://doi.org/10.1016/j.jrp.2010.03.003

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

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*(2), 197–210. https://doi.org/10.1177/00131649921969802

Wetzel, E., & Carstensen, C. H. (2017). Multidimensional modeling of traits and response styles. *European Journal of Psychological Assessment, 33*(5), 352–364. https://doi.org/10.1027/1015-5759/a000291

Wetzel, E., Lüdtke, O., Zettler, I., & Böhnke, J. R. (2016). The stability of extreme response style and acquiescence over 8 years. *Assessment, 23*(3), 279–291. https://doi.org/10.1177/1073191115583714

Wood, J. K., Anglim, J., & Horwood, S. (2022). A less evaluative measure of Big Five personality: Comparison of structure and criterion validity. *European Journal of Personality, 36*(5), 809–824. https://doi.org/10.1177/08902070211012920

Wood, J. K., Anglim, J., & Horwood, S. (2023). Less evaluative measures of personality in job applicant contexts: The effect on socially desirable responding and criterion validity. *Journal of Personality Assessment*. https://doi.org/10.1080/00223891.2023.2251158. Advance online publication

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*(2), 168–190. https://doi.org/10.1177/1094428104263674

Ziegler, M., & Bühner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69*(4), 548–565. https://doi.org/10.1177/0013164408324469

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). Faking: Knowns, unknowns, and points of contention. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality* *assessment* (pp. 3–16). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0011

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods, 18*(4), 679–703. https://doi.org/10.1177/1094428115574518

Original Research Article

Educational and Psychological Measurement
2025, Vol. 85(6) 1237–1277

sagepub.com/journals-permissions
DOI: 10.1177/00131644251341843
journals.sagepub.com/home/epm
Sage

Disentangling Qualitatively Different Faking Strategies in High-Stakes Personality Assessments: A Mixture Extension of the Multidimensional Nominal Response Model

Timo Seitz[1], Ö. Emre C. Alagöz[1] and Thorsten Meiser[1]

Abstract

High-stakes personality assessments are often compromised by faking, where test-takers distort their responses according to social desirability. Many previous models have accounted for faking by modeling an additional latent dimension that quantifies each test-taker's degree of faking. Such models assume a homogeneous response strategy among all test-takers, reflected in a measurement model in which substantive traits and faking jointly influence item responses. However, such a model will be misspecified if, for some test-takers, item responding is only a function of substantive traits or only a function of faking. To address this limitation, we propose a mixture modeling extension of the multidimensional nominal response model (M-MNRM) that can be used to account for qualitatively different response strategies and to model relationships of strategy use with external variables. In a simulation study, the M-MNRM exhibited good parameter recovery and high classification accuracy across multiple conditions. Analyses of three empirical high-stakes datasets provided evidence for the consistent presence of the specified latent classes in different personnel selection contexts, emphasizing the importance of accounting for such kind of response behavior heterogeneity in high-stakes assessment data. We end the article with a discussion of the model's utility for psychological measurement.

[1] University of Mannheim, Mannheim, Germany

Corresponding Author:
Timo Seitz, Department of Psychology, University of Mannheim, L13,15-17–room 515, 68161 Mannheim, Germany.
Email: timo.seitz@uni-mannheim.de

Keywords

faking, mixture model, response strategies, high-stakes assessments, multidimensional item response theory

Introduction

Self-report personality questionnaires are frequently employed in high-stakes assessments like personnel selection (Diekmann & König, 2015; Nikolaou & Foti, 2018), as personality measures derived from self-report questionnaires have been found to predict job performance and other outcomes in various contexts (e.g., Ones et al., 2007; Sackett & Walmsley, 2014). However, considering that personality tests in high-stakes assessments carry important consequences for test-takers, there is the threat that test-takers deliberately present themselves in an overly favorable manner, that is, engage in faking. Many studies over the past few decades have shown that faking has several adverse effects on the psychometric properties of a test (Ziegler et al., 2011), including elevated mean scores (e.g., Birkeland et al., 2006), inflated correlations between trait scales (e.g., Christiansen et al., 2021), and biased rank orders of test-takers which ultimately alter selection decisions (e.g., Mueller-Hanson et al., 2003).

To account for the response bias of faking, several latent variable models have been developed to capture variance in item responses that is due to faking (e.g., Hendy et al., 2021; Seitz et al., 2024, 2025; Ziegler et al., 2015). These models typically assume a homogeneous response strategy among test-takers, reflected in a measurement model with a continuous latent faking dimension on which test-takers vary quantitatively. However, research has shown that test-takers in high-stakes assessments in fact employ qualitatively different response strategies (Griffith & Converse, 2011; Robie et al., 2007). In this case, a single measurement model does not fully capture the faking process and potentially yields biased estimates of person and item parameters.

In the present work, we address this limitation by extending the faking model by Seitz et al. (2024, 2025) in a mixture modeling framework. Throughout this article, we investigate the extended model in a simulation study and in a set of empirical datasets from three job application settings. Before presenting details about the model extension, we will first introduce previous model-based faking accounts as well as previous mixture modeling approaches.

Model-Based Approaches to Accounting for Faking

As mentioned above, there are multiple approaches that account for faking using latent variable modeling. One prominent approach is to use structural equation modeling and model responses from a personality inventory with a bifactor model (Hendy et al., 2021; see also Klehe et al., 2012; Schmit & Ryan, 1993). In such a model, all items load on a common general factor, which captures variance among the items of different trait scales and thus reflects an ‘‘ideal-employee’’ or faking factor. The items of the different trait scales additionally load on specific factors, which reflect the respective substantive personality traits after controlling for the general (faking) factor. Previous work fitting a bifactor model to personality data has found that substantive trait estimates based on scores of specific factors are less distorted by faking than trait estimates based on classical scale scores (Hendy et al., 2021), and that the general faking factor is related to external covariates (Klehe et al., 2012).

Another approach to accounting for faking is to apply multidimensional item response theory (IRT). Multidimensional IRT models have frequently been used to account for response styles in rating scale data by specifying additional latent dimensions that reflect the response styles of interest (see Henninger & Meiser, 2020, for an overview). In many cases, these models are applications of the multidimensional nominal response model (MNRM; Takane & de Leeuw, 1987). In the parameterization of the model by Falk and Cai (2016; see also Thissen & Cai, 2016), item responses are modeled through the following softmax function which accounts for the influence of D latent dimensions:

$$p(Y_{ni}=k \mid \boldsymbol{\alpha}_i,\ \boldsymbol{S}_i,\ \boldsymbol{\gamma}_i,\ \boldsymbol{\theta}_n) = \frac{\exp((\boldsymbol{\alpha}_i \circ \boldsymbol{s}_{ik})' \boldsymbol{\theta}_n + \gamma_{ik})}{\sum_{m=0}^{K} \exp((\boldsymbol{\alpha}_i \circ \boldsymbol{s}_{im})' \boldsymbol{\theta}_n + \gamma_{im})}. \tag{1}$$

$Y \in \{0, 1, \ldots, k, \ldots, K\}$ is a random variable representing an item response, with k denoting its realization (i.e., the selected response category). The probability of response category k selected by person n on item i depends on a vector of person parameters $\boldsymbol{\theta}_n$ (containing trait scores of person n on the D dimensions), a vector of item-category intercepts $\boldsymbol{\gamma}_i$ (containing intercept values of item i for the $K+1$ categories), a vector of item slopes $\boldsymbol{\alpha}_i$ (containing factor loadings [aka discrimination parameters] of item i with respect to the D dimensions), and a matrix of scoring weights $\boldsymbol{S}_i$. Scoring weights reflect the relationship between category k and dimension d on item i, and can hence be used to specify the latent dimensions that should be modeled. For modeling substantive traits, a scoring weight vector of evenly-spaced integer values is typically specified, following the partial credit model (Masters, 1982) for ordinal responses as a special case of the nominal response model (Bock, 1972; Thissen & Steinberg, 1986). For response styles such as extreme response style (ERS) or midscale response style (MRS), a 0/1 scoring scheme is usually employed, where categories triggered by high levels of the respective response style are coded as 1 and other categories are coded as 0.

Seitz et al. (2024, 2025) transferred the method of specifying scoring weights to the response bias of faking. In particular, they set scoring weights of a faking dimension to values representing the desirability of a response category on a given item. When modeling responses to a personality test designed to measure three substantive traits with a 7-point Likert scale, the scoring weight matrix $\boldsymbol{S}_i$ for an item measuring the first substantive trait can be denoted as:

$$\boldsymbol{S}_i = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \text{des}_{i0} & \text{des}_{i1} & \text{des}_{i2} & \text{des}_{i3} & \text{des}_{i4} & \text{des}_{i5} & \text{des}_{i6} \end{pmatrix}, \tag{2}$$

where scoring weights of the substantive trait dimensions not measured by the particular item are set to 0 and des_{ik} stands for category k's desirability on item i. To get the desirability values in an empirical setting, one can conduct a pilot study in which participants are instructed to rate the desirability of each category of each item with respect to the social context of the actual personality testing (typically with respect to an application for a particular job; Seitz et al., 2025). Usually, items vary in how desirable the respective categories are, and the relationship between categories and desirability is oftentimes not strictly monotonic (Kuncel & Tellegen, 2009). Thus, when specifying the faking dimension as described, item-specific and potentially nonmonotonic faking effects can be modeled. In previous work, this way of modeling faking has been shown to significantly improve model fit, debias inflated correlations between substantive traits (Seitz et al., 2025), and enhance the estimation of individual substantive trait scores (Seitz et al., 2024).

Mixture Modeling Approaches to Accounting for Heterogeneity in Response Behavior

Independent of the modeling of faking, there are other methods to account for different kinds of heterogeneity in response behavior. One parametric approach is to use mixture modeling. These models assume that the data consists of distinct, unobserved subpopulations, so-called latent classes, each of which is associated with a separate measurement model (e.g., McLachlan & Peel, 2000; von Davier & Rost, 2006). That is, parameter values or the entire model structure may not be constant for all test-takers but vary between classes. Substantively, the different classes are assumed to represent distinct response processes, varying response strategies, or other types of heterogeneity. To account for such heterogeneous response behavior, separate measurement models are estimated based on assigning test-takers to the different classes. As a result, mixture models yield estimates of class proportions and probabilities of class membership for each test-taker.

In psychometrics, mixture models have been extensively used to account for heterogeneity in item responding. Many studies have followed an exploratory approach. For example, mixture-distribution IRT models have been applied to study heterogeneity in the use of rating scales (Rost, 1991). Such analyses have often found two classes characterized by varying threshold distances, typically treated as manifestations of ERS (e.g., Böckenholt & Meiser, 2017; Eid & Rauber, 2000; Gollwitzer et al., 2005). However, the interpretation of classes in exploratory mixture models is usually post hoc, atheoretical, and can become cumbersome in the case of complex model structures.

As opposed to fully exploratory mixture models, there are confirmatory mixture modeling approaches. These models impose theoretically motivated parameter constraints to account for specific forms of heterogeneity. A prominent example of such models are HYBRID models (Yamamoto, 1987, 1989), which assume two classes with model structures related to different cognitive processes. One class is specified in terms of a regular IRT model, whereas a model of stochastic independence is specified in the second class (see von Davier & Yamamoto, 2004; Yamamoto & Everson, 1995; for extensions). Thus, the HYBRID model separates regular test-takers from test-takers who choose response categories randomly. An overview of mixture-distribution IRT and HYBRID models can be found in von Davier and Yamamoto (2007).

Furthermore, other confirmatory mixture models have been developed in recent years to account for heterogeneity in rating scale responses. Tijmstra et al. (2018) proposed a two-class mixture IRT model that models a qualitatively different use of the midpoint category of a rating scale. Similar to a HYBRID model, a regular (in this case, ordinal) IRT model is specified in one class, whereas the second class is parameterized in terms of an item response tree (IRTree) model where test-takers first decide whether to choose the midpoint category and then, given they have not chosen the midpoint category, indicate their actual endorsement level. The model thus separates test-takers who use the midpoint category as part of the ordinal scale for item endorsement from test-takers who treat the midpoint category as sort of a non-response. Similarly, Kim and Bolt (2021) developed a two-class mixture model that accounts for differences in how test-takers come to select extreme response categories. In this model, one class is specified in which the selection of extreme categories is determined by ERS, whereas the substantive trait influences the choice of extreme categories in the second class. That is, the model allows for interindividual differences regarding which latent dimensions affect item responses. Finally, Alagöz and Meiser (2024) proposed a four-class model in which the mixture components reflect a different use of ERS and MRS. To specify the classes, slopes of response style dimensions that are not used in a class are fixed to 0, leading to an ''ERS-only class,'' ''MRS-only class,'' ''ERS&MRS class,'' and ''no response style class.'' The model hence aims to detect fine-grained heterogeneity in test-takers' response style usage. In data applications of these three mixture models, all classes turned out to be empirically prevalent, providing evidence for the existence of heterogeneous response strategies in questionnaire data.

Heterogeneity of Response Strategies in High-Stakes Assessments

As noted above, previous faking models typically account for faking in terms of a latent variable that captures each test-taker's faking degree. That is, a single measurement model is specified in which item responding is a function of test-takers' substantive trait *and* faking levels. However, as for response styles, there is evidence that test-takers in high-stakes assessments do not only differ quantitatively in faking but also qualitatively. Evidence for this claim comes, for instance, from studies examining the prevalence of faking. These studies have come to the conclusion that many test-takers in high-stakes assessments do engage in faking but also that a considerable proportion do not show self-presentational behavior (see Griffith & Converse, 2011; for an overview). Griffith et al. (2007), for example, retested job applicants under anonymous conditions and observed that 30% to 50% of applicants had significantly elevated their scores in the preceding application whereas the remaining applicants had not (see also Arthur et al., 2010). Similarly, using the randomized response technique, König et al. (2011) found that 32% of applicants in the United States exaggerate positive features in job application contexts whereas the other 68% do not (see also Donovan et al., 2003). Furthermore, evidence for qualitative differences in test-takers' faking behavior is provided by studies investigating the thought process in high-stakes personality testings (Robie et al., 2007; Röhner et al., 2025). Robie et al. (2007), for example, asked test-takers to think aloud while they were responding to a personality questionnaire under high-stakes conditions. Based on an analysis of verbal protocols, they found three groups: a group of test-takers referring to themselves and the "ideal" applicant while responding, a group of test-takers only considering themselves, as well as a group of test-takers exclusively responding from the perspective of the "ideal" applicant.

Considering this line of research, it is questionable whether the response process associated with faking can be described by a single continuous faking variable and a homogeneous loading structure of substantive traits and faking across test-takers. If, for subsets of test-takers, item responding is only a function of substantive traits or, to the other extreme, only a function of faking, a model assuming a joint influence of substantive traits and faking for all test-takers will be misspecified. For test-takers only responding according to substantive traits, the model will be inappropriate because substantive trait scores will be adjusted for an estimated faking degree that has no actual foundation since faking has not influenced item responding. Likewise, for test-takers for whom item responding is only a function of faking, the model will be inappropriate because it will nonetheless yield substantive trait score estimates for these test-takers. Apart from an inappropriate estimation of person parameters, one can also expect that the estimation of item parameters and latent correlations will be biased if test-takers differ qualitatively in how they (do not) align responses with desirability (see the simulation below).

Mixture Multidimensional Nominal Response Model (M-MNRM)

To allow for different response strategies in the modeling of faking, we propose a mixture extension of the MNRM that Seitz et al. (2024, 2025) used to account for faking. We hereby follow a confirmatory mixture modeling approach by constraining class-specific model parameters based on the definition of classes (see Alagöz & Meiser, 2024, for a similar approach). In particular, we specify three latent classes reflecting the response strategies test-takers may use in high-stakes assessments (see the above-described study by Robie et al., 2007). The first class represents a response strategy where test-takers select categories based on a combination of substantive traits and a faking dimension (‘‘S&F class’’), which is equivalent to the measurement model in the non-mixture version of the MNRM. Such a response behavior is conceivable because the obvious goal of making a favorable impression conflicts with the goal of staying true to oneself (Kuncel et al., 2011), so test-takers may want to find a compromise. Also, the conflict between wanting the job and social norms like telling the truth may lead to the joint consideration of substantive traits and faking. The second class reflects a response strategy where test-takers only respond based on substantive traits while faking does not influence item responses (‘‘S-only class’’). Reasons for such a response behavior can be that some test-takers do not know how to portray themselves favorably and hence do not engage in faking (Marcus, 2009; Ziegler, 2011), that some are afraid of being detected as liar or imposter (Röhner et al., 2025; Turner, 2022), or that some deliberately want to be honest in order to avoid being selected for a job they do not fit to (Kanfer et al., 2001; Saks & Ashforth, 2002). The third class represents a response strategy where item responding is not determined by substantive traits but solely by a faking dimension (‘‘F-only class’’). Such a response behavior can occur when test-takers want to be hired at any cost and therefore only consider desirability aspects of the items, or when test-takers want to compensate for poor scores on other relevant selection criteria (such as cognitive ability tests or grades).

To implement the three classes in the mixture model, one can impose class-specific model constraints, namely, set slopes of latent dimensions that are not part of a given response strategy to 0 for the respective class. Technically, the model equation of the mixture MNRM (M-MNRM) can be written as:

$$p(Y_{ni}=k \mid \boldsymbol{\alpha}_{ic}, \boldsymbol{S}_i, \boldsymbol{\gamma}_{ic}, \boldsymbol{\theta}_n) = \sum_{c=1}^{3} \frac{\exp\left((\boldsymbol{\alpha}_{ic} \circ \boldsymbol{s}_{ik})' \boldsymbol{\theta}_n + \gamma_{ikc}\right)}{\sum_{m=0}^{K} \exp\left((\boldsymbol{\alpha}_{ic} \circ \boldsymbol{s}_{im})' \boldsymbol{\theta}_n + \gamma_{imc}\right)} p(\zeta_n = c) \,, \tag{3}$$

where $\zeta_n \in \{1, 2, 3\}$ denotes the class membership of person n. This equation describes the total probability of response k for person n on item i by multiplying the class-specific response probability with the probability of being a member of this class before summing across the three classes. The term $p(\zeta_n = c)$ is often referred to as the proportion of class c. When conditioning on person n's class membership, the equation boils down to the class-specific probability of response k for person n on item i:

$$p(Y_{ni}=k \mid \boldsymbol{\alpha}_{ic}, \boldsymbol{S}_i, \boldsymbol{\gamma}_{ic}, \boldsymbol{\theta}_n, \zeta_n = c) = \frac{\exp((\boldsymbol{\alpha}_{ic} \circ \boldsymbol{s}_{ik})' \boldsymbol{\theta}_n + \gamma_{ikc})}{\sum_{m=0}^{K} \exp((\boldsymbol{\alpha}_{ic} \circ \boldsymbol{s}_{im})' \boldsymbol{\theta}_n + \gamma_{imc})} \,. \tag{4}$$

Consider modeling faking in a personality questionnaire measuring three substantive traits with three items each, the slope matrices of the three latent classes in the M-MNRM can be denoted as:

$$\boldsymbol{A}_{c=1} = \begin{pmatrix} \alpha_{1\mathrm{S}_1} & 0 & 0 & \alpha_{1\mathrm{F}} \\ \alpha_{2\mathrm{S}_1} & 0 & 0 & \alpha_{2\mathrm{F}} \\ \alpha_{3\mathrm{S}_1} & 0 & 0 & \alpha_{3\mathrm{F}} \\ 0 & \alpha_{4\mathrm{S}_2} & 0 & \alpha_{4\mathrm{F}} \\ 0 & \alpha_{5\mathrm{S}_2} & 0 & \alpha_{5\mathrm{F}} \\ 0 & \alpha_{6\mathrm{S}_2} & 0 & \alpha_{6\mathrm{F}} \\ 0 & 0 & \alpha_{7\mathrm{S}_3} & \alpha_{7\mathrm{F}} \\ 0 & 0 & \alpha_{8\mathrm{S}_3} & \alpha_{8\mathrm{F}} \\ 0 & 0 & \alpha_{9\mathrm{S}_3} & \alpha_{9\mathrm{F}} \end{pmatrix} \text{ for the ''S\&F class,''} \tag{5}$$

$$\boldsymbol{A}_{c=2} = \begin{pmatrix} \alpha_{1\mathrm{S}_1} & 0 & 0 & 0 \\ \alpha_{2\mathrm{S}_1} & 0 & 0 & 0 \\ \alpha_{3\mathrm{S}_1} & 0 & 0 & 0 \\ 0 & \alpha_{4\mathrm{S}_2} & 0 & 0 \\ 0 & \alpha_{5\mathrm{S}_2} & 0 & 0 \\ 0 & \alpha_{6\mathrm{S}_2} & 0 & 0 \\ 0 & 0 & \alpha_{7\mathrm{S}_3} & 0 \\ 0 & 0 & \alpha_{8\mathrm{S}_3} & 0 \\ 0 & 0 & \alpha_{9\mathrm{S}_3} & 0 \end{pmatrix} \text{ for the ''S-only class,'' and} \tag{6}$$

$$\boldsymbol{A}_{c=3} = \begin{pmatrix} 0 & 0 & 0 & \alpha_{1\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{2\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{3\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{4\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{5\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{6\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{7\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{8\mathrm{F}} \\ 0 & 0 & 0 & \alpha_{9\mathrm{F}} \end{pmatrix} \text{ for the ''F-only class,''} \tag{7}$$

where the rows reflect the items and the columns reflect the latent dimensions.

Furthermore, to model relationships between the use of response strategies and external variables, class membership can be predicted by a set of covariates. This can be achieved through a latent multinomial logistic regression of class membership on the covariates:

$$\pi_{nc} = p(\zeta_n = c \mid \boldsymbol{X}_n = \boldsymbol{x}_n) = \frac{\exp(\beta_{0c} + \sum_{p=1}^{P} \beta_{pc} x_{np})}{\sum_{m=1}^{3} \exp(\beta_{0m} + \sum_{p=1}^{P} \beta_{pm} x_{np})} \ . \tag{8}$$

$\boldsymbol{X}$ is a multivariate random variable representing the P covariates, $\boldsymbol{x}_n$ denotes the realizations for person n. β_{0c} is the regression intercept for class c, β_{pc} are regression slopes that reflect the effect of covariate p on class c. In this article, the ''S&F class'' represents the reference class. Hence, regression coefficients pertaining to this class are fixed to 0, such that intercepts and slopes pertaining to the other classes are to be interpreted with respect to the ''S&F class.''

It is important to note that the model does *not* imply that test-takers in the ''S-only class'' per se do not have a faking person parameter or, correspondingly, that test-takers in the ''F-only class'' per se do not have substantive trait person parameters. Instead, the model assumes that test-takers in the ''S-only class'' (''F-only class'') do not consider the faking dimension (substantive trait dimensions) when responding to the items. Similarly, the proposed model is not equivalent to a non-mixture model in which ''S-only'' test-takers (''F-only'' test-takers) have faking person parameters (substantive trait person parameters) of 0. Whereas a person parameter of 0 simply reflects one possible value on the dimension's latent continuum (oftentimes representing the latent mean), an item slope of 0 implies that the dimension does not explain any variance in item responses, which captures the idea of qualitatively different response strategies (see also Alagöz & Meiser, 2024). More information on this matter, including a data illustration, can be found in the Supplemental Material.

Note also that, in our parameterization, non-fixed item slopes are class-invariant. This allows for measuring the same latent variables across classes, such that classes only differ concerning the loading structure of items and factors (Alagöz & Meiser, 2024; Kim & Bolt, 2021). Allowing non-fixed item slopes to vary freely would potentially change the meaning of the latent variables across classes. However, item-category intercepts are class-specific in our parameterization. This is because the three classes are likely to differ in their response distributions. As class-specific intercepts can capture different response distributions across classes, modeling intercepts in an unconstrained manner can be assumed to facilitate class separation when estimating the model.

Differences to Other Faking Mixture Models

Before coming to details about the estimation of the model, we will first delineate how the M-MNRM differs from other faking models including mixture components. Zickar et al. (2004) applied a mixture-distribution IRT model to personality data from an applicant sample. They found three classes characterized by a different ordering and spacing of threshold parameters, which they interpreted as an honest, slight-faking, and extreme-faking class. Nevertheless, since such a mixture modeling approach is fully exploratory, it remains uncertain whether the resulting classes truly capture faking. It may well be that the classes in fact represent other response tendencies. Also, such a model conceptualizes faking only as a discrete variable. Related work, however, emphasized the continuous nature of faking (Ziegler et al., 2015).

Böckenholt (2014; see also Leng et al., 2020) modeled test-takers' misreporting behavior in sensitive survey questions. This model specifies for every item a binary latent class variable indicating whether a test-taker edits his or her retrieved response. If a test-taker has decided to edit, the selection of desirable response categories is modeled by a transition function. Even though the model is conceptually appealing, it comes with the limitation that only one substantive trait can be modeled. Also, the selection of categories when editing occurs is modeled in a monotonic way, such that the selected categories are assumed to be always higher (lower) if the substantive trait is generally desirable (undesirable).

Brown and Böckenholt (2022) developed a grade-of-membership model to account for intermittent faking. This model assumes that each item response either stems from a ''real'' (honest) or ''ideal'' (faking) class that is predicted by a latent editing factor and item characteristics. Within each class, item responses are either a function of substantive traits or a function of a faking factor. Though theoretically elaborate, the model does not include a class where responses are influenced by both substantive traits and faking, which is conceivable considering the findings by Robie et al. (2007) described above. Also, like the approaches by Zickar et al. (2004) and Böckenholt (2014), the model does not explicitly account for nonmonotonic faking effects in a way that, for some items, high faking levels make the selection of non-extreme categories more likely (see Kuncel & Tellegen, 2009; Seitz et al., 2024, 2025).

Model Estimation

The M-MNRN can be estimated in a Bayesian Markov chain Monte Carlo (MCMC) procedure. Therefore, we implemented the full model (Equation 3), in which the latent regression of class membership on covariates (Equation 8) is nested, in the program *JAGS* (version 4.3.1; Plummer, 2017). We accessed JAGS via the *R* environment (version 4.4.0) using the package *runjags* (Denwood, 2016), and employed the packages *coda* (Plummer et al., 2006) and *MCMCvis* (Youngflesh, 2018) for processing MCMC outputs. The JAGS syntax and R code for estimating the model can be found at https://osf.io/vwqf3/.

For model estimation, the following priors were used: Non-fixed item slopes were sampled from a positively-truncated normal distribution ($\alpha_{id} \sim N^{+}(0,\ 2^2)$). Class-specific item-category intercepts were drawn from an uncensored normal distribution ($\gamma_{ikc} \sim N(0,\ 5^2)$), with the intercept of the first category fixed to 0 due to model identification. Substantive trait and faking scores were drawn from a multivariate normal distribution ($\boldsymbol{\theta}_n \sim MVN(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$), where $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ was a variance-covariance matrix with unit variances. Thus, covariances represented correlations, which had a uniform prior distribution ($\rho_{dd'} \sim U(-1,\ 1)$). Latent regression coefficients were sampled from a normal distribution ($\beta_{0c}, \beta_{pc} \sim N(0,\ 2^2)$), with coefficients pertaining to the ''S&F class'' fixed to 0. Class membership was drawn from a categorical distribution ($\zeta_n \sim Cat(\boldsymbol{\pi}_n)$), in which $\boldsymbol{\pi}_n$ was a vector of person-specific class probabilities resulting from the latent regression of class membership at the respective MCMC iteration.

To obtain point estimates of continuous model parameters, means of posterior distributions were computed. For class membership as a discrete model parameter, the posterior mode was considered (i.e., modal assignment; Dias & Vermunt, 2008).

Because of the latent regression of class membership in the M-MNRM, the model does not include explicit class proportion parameters. However, mean class probabilities across persons and MCMC iterations can be calculated to get class proportion estimates. Note that, in the present model, class labels are not arbitrary because different measurement models are explicated for the three classes. This prevents the problem of label switching (Stephens, 2000) that is frequently encountered in the Bayesian estimation of mixture models.

Simulation Study

To investigate the M-MNRM under different class proportion conditions, we conducted a simulation study analyzing parameter recovery and the model's superiority over non-mixture models. Also, the simulation should examine whether the correct model (non-mixture vs. mixture) is selected when either one or multiple classes are present in the data.

Simulation Conditions

The simulation featured 10 class proportion conditions: In Condition 1, class proportions were (33.3%, 33.3%, 33.3%), that is, the three classes were equally sized. Conditions 2 to 4 were conditions in which either the ''S&F class'' (Condition 2: (60%, 20%, 20%)), the ''S-only class'' (Condition 3: (20%, 60%, 20%)), or the ''F-only class'' (Condition 4: (20%, 20%, 60%)) was dominant. In Conditions 5 to 7, one class was absent, either the ''S&F class'' (Condition 5: (0%, 50%, 50%)), the ''S-only class'' (Condition 6: (50%, 0%, 50%)), or the ''F-only class'' (Condition 7: (50%, 50%, 0%)). Conditions 8 to 10 represented data situations with non-mixture populations. That is, only one class was present in the data, either the ''S&F class'' (Condition 8: (100%, 0%, 0%)), the ''S-only class'' (Condition 9: (0%, 100%, 0%)), or the ''F-only class'' (Condition 10: (0%, 0%, 100%)).

Data Generation and Fitted Models

For every simulation condition, we simulated item responses of a questionnaire measuring 3 substantive traits with 10 items each on a 7-point Likert scale. We chose this simulation design to examine data situations representative of empirical high-stakes datasets (see the empirical demonstration below). To generate the data, we took the following steps:

1. Item slopes α_{id}: Item slopes of substantive traits and faking were sampled from $U(0.5,\ 1)$.
2. Scoring weights s_{idk}: Scoring weight vectors of substantive traits were set to $(0\quad 1\quad 2\quad 3\quad 4\quad 5\quad 6)$. Scoring weight vectors of faking also had a range

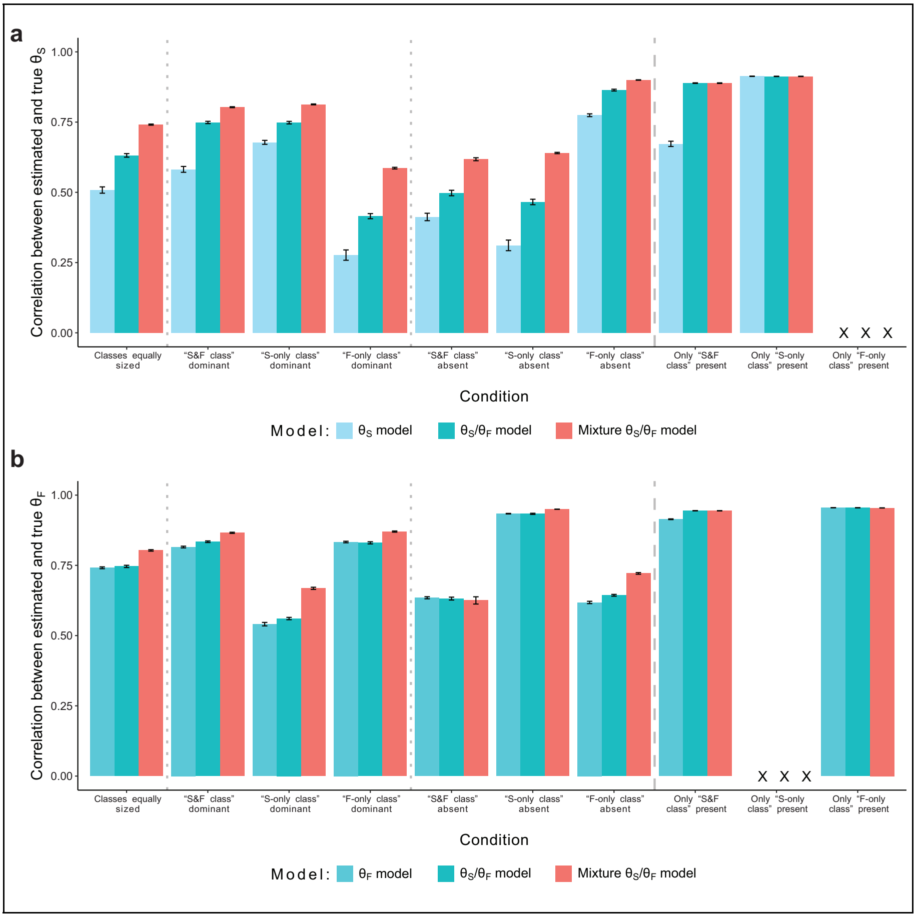


Figure 1. Simulation Study: Recovery of Substantive Trait and Faking Scores. (a) Recovery of Substantive Trait Scores and (b) Recovery of Faking Scores.

Note. Values reflect the mean correlations (using Fisher's *z*-transformation) between estimated and true substantive trait scores (a) or faking scores (b) across replications within a condition. Results for substantive traits are aggregated across the three substantive traits used in the simulation. Error bars represent the standard error of the mean. "X" denotes that a proper recovery of the particular parameters is precluded in the respective condition because they have not influenced item responses in the data generation.

from 0 to 6 but varied between items to emulate a situation in which relationships between response categories and desirability are item-specific. Specifically, within every substantive trait scale, scoring weight vectors of faking were generated reflecting monotonically increasing, nonmonotonically increasing, as well as inverted-U-shaped relations between categories and desirability (see Figure 1 in Seitz et al., 2024, for details).

3. Item-category intercepts γ_{ikc}: For every item, the intercept of the first category was fixed to 0. Intercepts of the remaining categories were simulated by sampling sorted thresholds τ_{ikc} from $U(-2,\ 2)$ before the resulting values were transformed into cumulative thresholds representing intercepts: $\gamma_{ikc} = -\sum_{m=0}^{k} \tau_{imc}$. This procedure was carried out independently for the three classes.
4. Substantive trait and faking scores θ_{nd}: For each of $N = 1,500$ simulated test-takers, three substantive trait scores and a faking score were drawn from $MVN(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & & & \\ \mathit{0} & 1 & & \\ \mathit{.30} & \mathit{-.30} & 1 & \\ 0 & .30 & -.30 & 1 \end{pmatrix}$. The assignment of the latent correlations printed in italics to the three substantive trait pairs was randomized between replications.
5. Latent regression coefficients β_{0c}, β_{pc}: Three covariates of class membership were considered, one with a null effect, one with weak effects, and one with strong effects. With the ''S&F class'' as the reference class, latent regression coefficients pertaining to this class were fixed to 0. Regression slopes of the null-effect covariate pertaining to the remaining classes were also set to 0, whereas slopes of the weak-effects and strong-effects covariates were sampled from $N(1,\ 0.2^2)$ and $N(2,\ 0.2^2)$, respectively. Half of the regression slopes within each replication were multiplied by -1 to simulate both positive and negative covariate effects. Regression intercepts were specified to lead to the class proportions of the respective simulation condition.
6. Covariate values X_{np}: For all simulated test-takers, values on the three covariates were drawn from $MVN(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ as a diagonal matrix with unit variances.
7. Class membership ζ_n: Using Equation (8), class probabilities $\boldsymbol{\pi}_n$ were computed for every simulated test-taker, which were then used to sample the actual class membership from $Cat(\boldsymbol{\pi}_n)$.[1]
8. Based on the generated item parameters, substantive trait and faking scores, as well as class memberships, item responses were simulated using Equation (4).
9. Steps 1 to 8 were replicated such that there were 30 generated datasets per condition.

The data generation was performed in R using the packages *MASS* (Venables & Ripley, 2002) and *mirt* (Chalmers, 2012). Within each condition, four models were fitted to all generated datasets: a model only accounting for a faking dimension with specified scoring weights (''θ_F model''), a model only accounting for substantive traits (''θ_S model''), a model accounting for substantive traits and faking (''θ_S/θ_F model''), as well as the M-MNRM accounting for substantive traits and faking with the three latent classes (''mixture θ_S/θ_F model''). Scoring weights were specified as in the data generation. The three non-mixture models were also estimated using the Bayesian estimation framework described above (the non-mixture model syntaxes are also available at https://osf.io/vwqf3/). Four parallel MCMC chains were run for every model. The estimation featured a burnin phase of 2,000 iterations, followed by 5,000 regular iterations.

Simulation Results

To check model convergence, we considered $\hat{R}$ values of continuous model parameters (Gelman & Rubin, 1992). These were below 1.1 for all models. Also, we visually inspected MCMC chains of item parameters, latent correlations, and latent regression coefficients and found that trace plots were well mixed.

Model Selection. As mentioned above, we examined if model selection criteria correctly chose the mixture model when there was more than one class in the data, and, crucially, if they chose a non-mixture model when the data came from a non-mixture population. Therefore, we considered the deviance information criterion (DIC; Spiegelhalter et al., 2002) as calculated by Gelman et al. (2004), the widely applicable information criterion (WAIC; Watanabe, 2010), and the leave-one-out information criterion (LOOIC; Vehtari et al., 2017). These measures balance model fit and model parsimony by penalizing the model's mere fit to the data with the effective number of parameters or by considering out-of-sample predictive accuracy. Table 1 shows the percentages with which DIC, WAIC, and LOOIC selected the correct data-generating model. In conditions with more than one class present in the data, every model selection criterion correctly selected the "mixture θ_S/θ_F model" in all replications. In conditions with only one class in the data, the performance of model selection criteria was still high (above 80%), though in some cases incorrect models were selected. However, incorrect model selections were not only due to an overselection of the mixture model but also due to false model selections within the three non-mixture models. Percentages of incorrect mixture model selections in conditions with only one class were below 10%. It is also important to note that, in replications in which the mixture model was falsely selected, the number of simulated test-takers assigned to a truly non-existent class was negligible (see hit rates below).

Parameter Recovery. To evaluate parameter recovery, we considered the bias of estimation to investigate if parameters were systematically over- or underestimated as well as the root mean square error (RMSE) to investigate the accuracy of estimation. For the recovery of substantive trait and faking scores, we considered the correlation between estimated and true parameters. To examine how well the individual class membership was recovered, we considered the hit rate, which indicates the percentage of persons correctly assigned to their respective class and thus reflects a measure of classification accuracy.

Table 1. Simulation Study: Percentages of Correctly Selected Models.

Condition	Model selection criterion		
	DIC	WAIC	LOOIC
Classes equally sized	100.0% (100.0%)	100.0% (100.0%)	100.0% (100.0%)
"S&F class" dominant	100.0% (100.0%)	100.0% (100.0%)	100.0% (100.0%)
"S-only class" dominant	100.0% (100.0%)	100.0% (100.0%)	100.0% (100.0%)
"F-only class" dominant	100.0% (100.0%)	100.0% (100.0%)	100.0% (100.0%)
"S&F class" absent	100.0% (100.0%)	100.0% (100.0%)	100.0% (100.0%)
"S-only class" absent	100.0% (100.0%)	100.0% (100.0%)	100.0% (100.0%)
"F-only class" absent	100.0% (100.0%)	100.0% (100.0%)	100.0% (100.0%)
Only "S&F class" present	80.0% (90.0%)	96.7% (100.0%)	90.0% (90.0%)
Only "S-only class" present	83.3% (93.3%)	96.7% (96.7%)	93.3% (96.7%)
Only "F-only class" present	86.7% (93.3%)	93.3% (100.0%)	86.7% (93.3%)

Note. Percentages are based on 30 replications per condition. In simulation conditions in which more than one class was present, the "mixture θ_S/θ_F model" was the underlying population model, whereas either the "θ_S/θ_F model," "θ_S model," or "θ_F model" was the population model in conditions in which only the respective class was present. Values in brackets reflect percentages of correct decisions concerning the question of whether a mixture or non-mixture model was the data-generating model. DIC = deviance information criterion; WAIC = widely applicable information criterion; LOOIC = leave-one-out information criterion.

Class proportions and class membership. The recovery of class proportions in the "mixture θ_S/θ_F model" is displayed in Table 2. Across conditions, class proportions were estimated with negligible bias and small RMSE. Bias and RMSE did not systematically vary between conditions and the three classes.

Apart from an accurate estimation of overall class proportions, individual class membership was also recovered well, indicated by high hit rates (see Table 2). In conditions with more than one class in the data, hit rates ranged from 96.9% to 98.5%. In conditions with only one class, hit rates were close to 100%. That is, virtually no simulated test-taker was assigned to a class that was empty in the data generation, which gives another indication along with model selection criteria that a less complex (i.e., a non-mixture) model should be used in this case (see also the discussion below).

Latent regression coefficients. Similar to the recovery of class proportions, the "mixture θ_S/θ_F model" estimated latent regression coefficients without systematic bias (see Table 3). This applied to regression intercepts as well as regression slopes representing the covariate effects. RMSE was also small for both intercepts and slopes. Concerning intercepts, RMSE was relatively more pronounced in conditions with unequally sized classes compared to conditions with equal class proportions and conditions with one class being absent. Concerning slopes, RMSE increased slightly with stronger covariate effect sizes.

Substantive trait and faking scores. Figure 1a shows the recovery of substantive trait scores in models including substantive trait dimensions. The overall level of

Table 2. Simulation Study: Recovery of Class Proportions and Hit Rates.

Condition	Class					
	"S&F class"		"S-only class"		"F-only class"	
	Bias (RMSE)	HR	Bias (RMSE)	HR	Bias (RMSE)	HR
Classes equally sized	−0.001 (0.004)	0.975	0.000 (0.002)	0.981	0.000 (0.002)	0.979
"S&F class" dominant	0.000 (0.003)	0.981	0.000 (0.002)	0.981	0.000 (0.002)	0.973
"S-only class" dominant	0.000 (0.002)	0.973	0.000 (0.002)	0.984	0.000 (0.002)	0.978
"F-only class" dominant	−0.001 (0.003)	0.969	0.000 (0.002)	0.982	0.000 (0.003)	0.983
"S&F class" absent	0.002 (0.005)		−0.001 (0.002)	0.985	−0.001 (0.002)	0.984
"S-only class" absent	−0.002 (0.005)	0.980	0.002 (0.004)		0.000 (0.003)	0.982
"F-only class" absent	−0.002 (0.004)	0.981	0.000 (0.002)	0.982	0.002 (0.004)	
Only "S&F class" present	−0.004 (0.009)	0.998	0.002 (0.004)		0.002 (0.005)	
Only "S-only class" present	0.002 (0.004)		−0.003 (0.006)	0.997	0.001 (0.003)	
Only "F-only class" present	0.002 (0.003)		0.001 (0.003)		−0.003 (0.006)	0.998

Note. Values reflect the mean bias and RMSE (in brackets) of estimated class proportions across replications within a condition, as well as HRs. HRs are the mean percentages of simulated test-takers correctly assigned to their respective class. HR = hit rate; RMSE = root mean square error.

Table 3. Simulation Study: Recovery of Latent Regression Coefficients.

Condition	Bias (RMSE) of regression coefficients			
	Intercepts	Slopes of null-effect covariate	Slopes of weak-effects covariate	Slopes of strong-effects covariate
Classes equally sized	0.00 (0.09)	0.00 (0.07)	−0.03 (0.09)	0.02 (0.13)
"S&F class" dominant	−0.01 (0.11)	0.01 (0.08)	0.00 (0.10)	0.01 (0.11)
"S-only class" dominant	−0.02 (0.15)	−0.01 (0.07)	−0.01 (0.11)	0.03 (0.14)
"F-only class" dominant	0.00 (0.13)	0.01 (0.08)	0.00 (0.09)	−0.02 (0.13)
"S&F class" absent	−0.01 (0.06)	0.02 (0.07)	−0.03 (0.11)	0.03 (0.18)
"S-only class" absent	0.03 (0.06)	0.01 (0.06)	0.01 (0.07)	0.00 (0.09)
"F-only class" absent	0.00 (0.05)	0.01 (0.06)	0.02 (0.07)	0.01 (0.08)

Note. Values reflect the mean bias and RMSE (in brackets) of estimated latent regression coefficients across replications within a condition. In the condition in which the "S&F class" was absent, the "S-only class" is treated as the reference class. Results for conditions with only one class in the data are left out because class membership is a constant in these conditions, which precludes a proper recovery of regression coefficients. RMSE = root mean square error.

parameter recovery primarily varied with the proportion of simulated test-takers for whom item responding was not influenced by substantive traits. Within conditions, the "θ_S/θ_F model" generally exhibited higher correlations between estimated and true substantive trait scores than the "θ_S model." Crucially, the "mixture θ_S/θ_F model" improved correlations further in conditions in which more than one class was present in the data. In conditions with only one class, the "mixture θ_S/θ_F model" and "θ_S/θ_F model" did not differ.

A similar pattern emerged for the recovery of faking scores (see Figure 1b). The "mixture θ_S/θ_F model" yielded the highest correlations between estimated and true faking scores in most conditions with more than one class, whereas the "θ_S/θ_F model" and "θ_F model" differed only slightly. In the condition in which the "S&F class" was absent, all three models exhibited comparable levels of correlation. Again, in conditions with only one class, the "mixture θ_S/θ_F model" and "θ_S/θ_F model" performed equivalently.

Item slopes. The recovery of item slopes is presented in Figure 2. Substantive trait slopes (Figure 2a) were generally underestimated in the "θ_S model" and "θ_S/θ_F model" and had pronounced RMSE, unless the respective model matched the population model in conditions with only one class in the data. The "mixture θ_S/θ_F model" eliminated the bias in all conditions and reduced RMSE considerably. Likewise, estimates of faking slopes (Figure 2b) in the "θ_F model" and "θ_S/θ_F model" were negatively biased when there was more than one class in the data, whereas the "mixture θ_S/θ_F model" showed negligible bias. RMSE was also much smaller in the mixture model than in the non-mixture models. In conditions with only

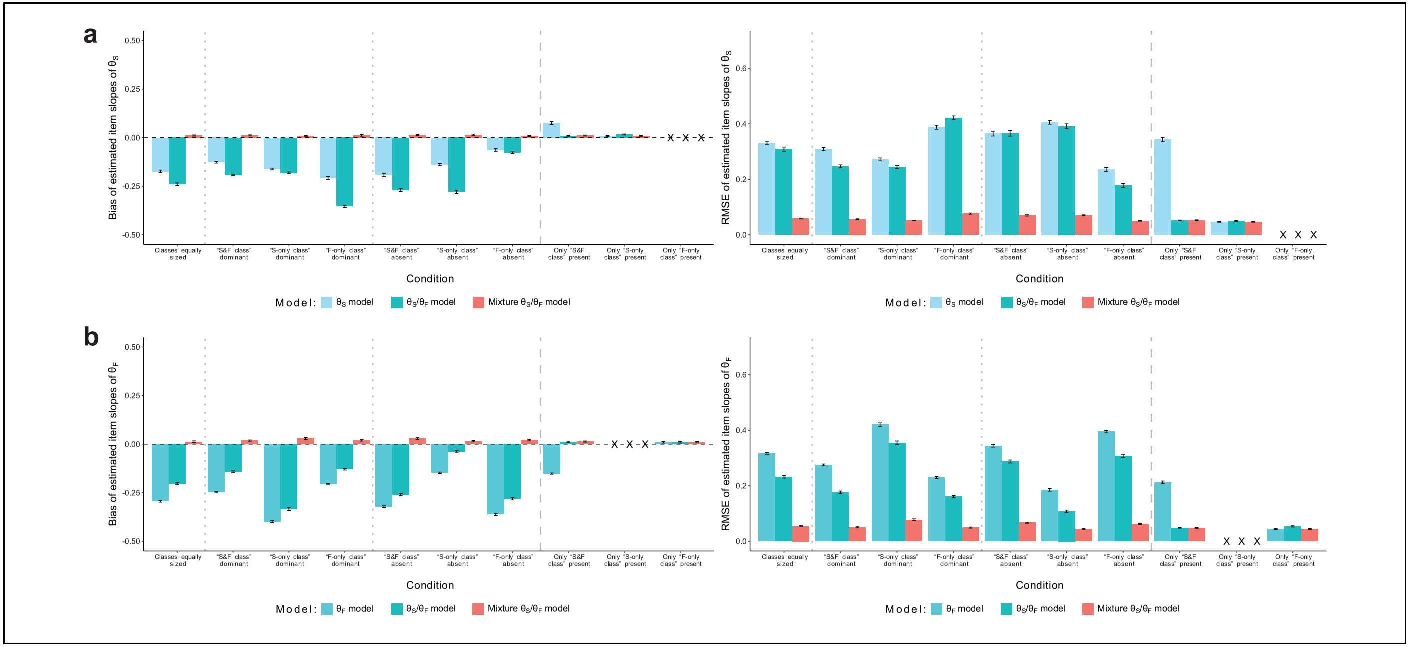


Figure 2. Simulation Study: Recovery of Item Slopes. (a) Recovery of Item Slopes of Substantive Traits and (b) Recovery of Item Slopes of Faking.

Note. Values reflect the mean bias or RMSE of estimated item slopes of substantive traits (a) or faking (b) across replications within a condition. Error bars represent the standard error of the mean. "X" denotes that a proper recovery of the particular parameters is precluded in the respective condition because they have not influenced item responses in the data generation. RMSE = root mean square error.

one class, the mixture model and the correctly specified non-mixture model both yielded unbiased estimates with equivalent RMSE.

Item-category intercepts. Figure 3 shows the recovery of item-category intercepts, which were class-specific in the data generation. The three non-mixture models generally yielded negatively biased intercept estimates with pronounced RMSE when the data consisted of more than one class. Only in conditions with just one class and if the respective model was not underparameterized, the non-mixture models estimated intercepts without bias and with small RMSE. In contrast, the mixture model's class-specific intercept estimates had very small to negligible bias and RMSE in all conditions.

Latent correlations. Latent correlations between substantive traits (see Figure 4a) were heavily overestimated in the ''θ_S model,'' especially in conditions with smaller proportions of simulated test-takers for whom item responding was influenced by substantive traits. The ''θ_S/θ_F model'' considerably reduced this bias, even yielding fully unbiased estimates in conditions with a small ''F-only class'' proportion. The ''mixture θ_S/θ_F model,'' however, estimated latent correlations between substantive traits without bias in all conditions. Also, RMSE was smaller in the mixture model than in the two non-mixture models. Regarding latent correlations between faking and substantive traits (see Figure 4b), the ''θ_S/θ_F model'' generally yielded underestimates with considerable RMSE, whereas the ''mixture θ_S/θ_F model'' afforded an unbiased estimation with smaller RMSE. In the condition in which only the ''S&F class'' was present, both models exhibited no bias and equivalent RMSE.

Empirical Demonstration

Along with the reported simulation study, we applied the M-MNRM to three empirical datasets from personnel selection contexts. This allowed us to examine the existence and prevalence of the three latent faking classes in high-stakes assessment data from different job application contexts.

Datasets

The datasets were made available by a Germany-based testing company that develops psychological tests for personnel selection. All three datasets contained responses from test-takers who had taken a personality test as part of an application for a job. Dataset 1 consisted of $N = 3,046$ test-takers applying for a bank apprenticeship at a financial institution in Germany (gender: 60.4% female, 39.6% male; age: $M = 18.22$ years, $SD = 1.98$, range $= [14, 29]$; this dataset was also analyzed by Seitz et al., 2025). Dataset 2 comprised $N = 1,824$ applicants for a police officer traineeship at a German police department (gender: 30.0% female, 70.0% male; age: $M = 21.02$ years, $SD = 4.61$, range $= [15, 39]$). Dataset 3 included $N = 2,007$ test-takers who had applied for a position as insurance agent at a Germany-based insurance company (gender: 28.7% female, 71.2% male).[2]

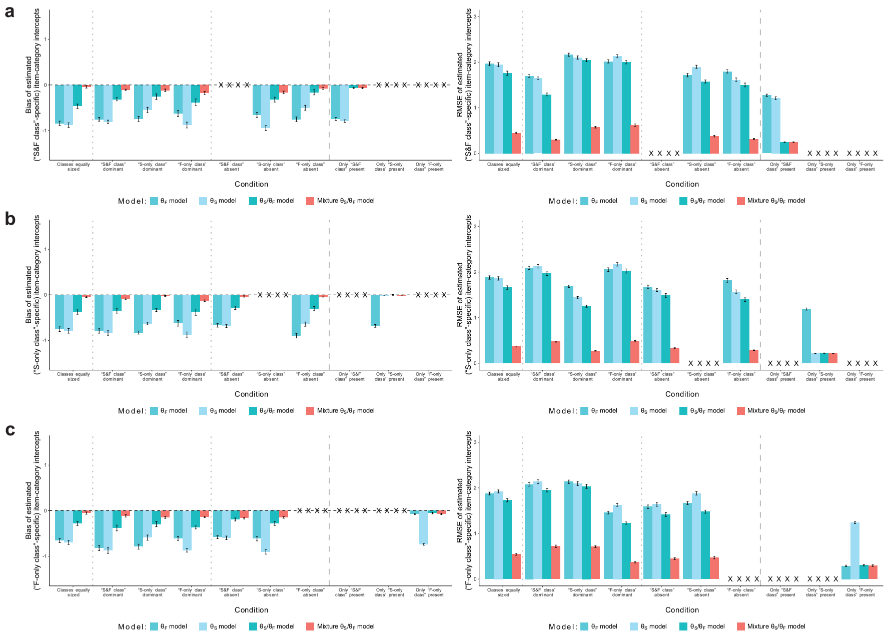


Figure 3. Simulation Study: Recovery of Item-Category Intercepts. (a) Recovery of Item-Category Intercepts of the "S&F Class"; (b) Recovery of Item-Category Intercepts of the "S-only Class"; and (c) Recovery of Item-Category Intercepts of the "F-only Class."

Note. Values reflect the mean bias or RMSE of estimated item-category intercepts with respect to the true values in the "S&F class" (a), "S-only class" (b), or "F-only class" (c) across replications within a condition. For the mixture model, the respective class-specific intercept estimates are considered. Error bars represent the standard error of the mean. "X" denotes that a proper recovery of the particular parameters is precluded in the respective condition because they have not influenced item responses in the data generation. RMSE = root mean square error.

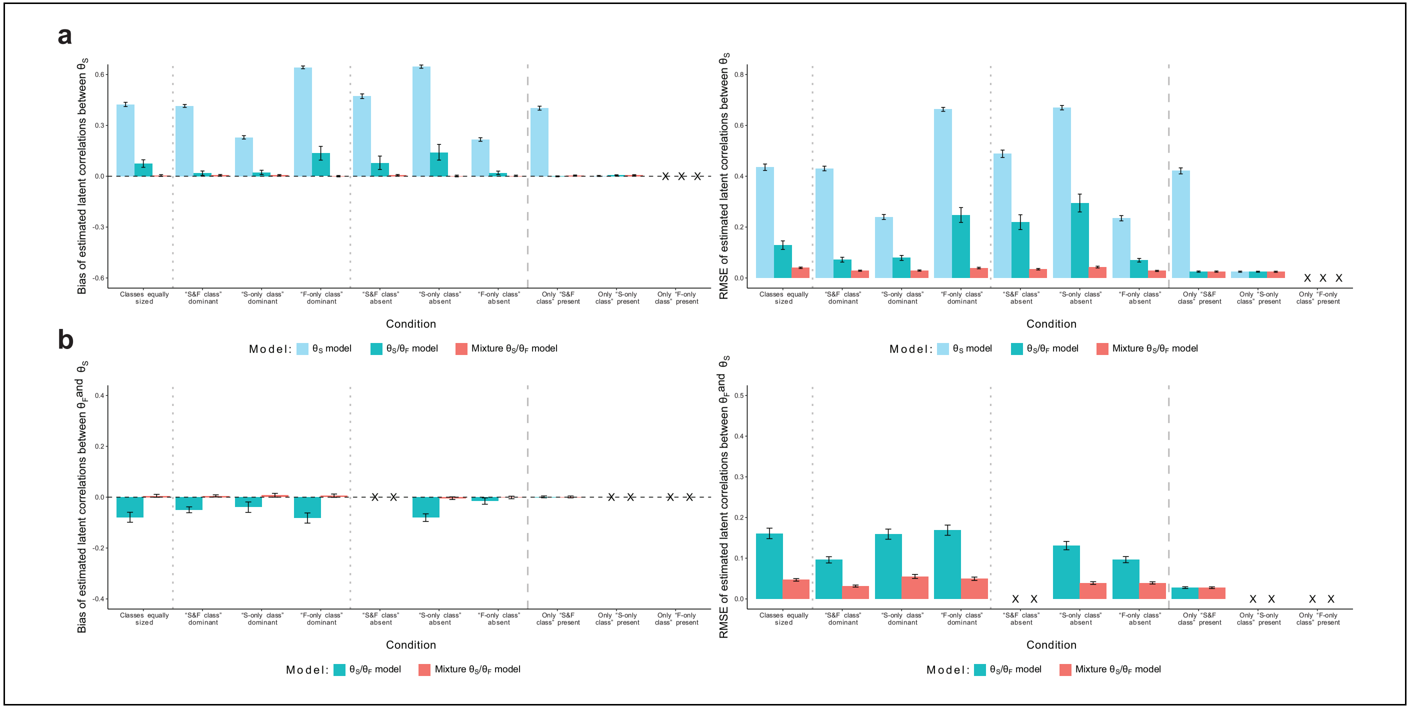


Figure 4. Simulation Study: Recovery of Latent Correlations. (a) Recovery of Latent Correlations Between Substantive Traits and (b) Recovery of Latent Correlations Between Faking and Substantive Traits.

Note. Values reflect the mean bias or RMSE of estimated latent correlations between substantive traits (a) or faking and substantive traits (b) across replications within a condition. Results are aggregated across the three substantive traits used in the simulation. Error bars represent the standard error of the mean. "X" denotes that a proper recovery of the particular parameters is precluded in the respective condition because they have not influenced item responses in the data generation. RMSE = root mean square error.

In the present empirical demonstration, we modeled data from three substantive trait scales that were available in all datasets. These scales assessed the personality traits of Emotional Stability, Extraversion, and Conscientiousness. Emotional Stability was measured with 12 items (Dataset 1: Cronbach's $\alpha = .75$; Dataset 2: $\alpha = .74$; Dataset 3: $\alpha = .70$), Extraversion with 9 items (Dataset 1: $\alpha = .74$; Dataset 2: $\alpha = .67$; Dataset 3: $\alpha = .72$), and Conscientiousness with 10 items (Dataset 1: $\alpha = .79$; Dataset 2: $\alpha = .77$; Dataset 3: $\alpha = .66$). Item responses were given on a 7-point Likert scale (0 = *does not apply at all* to 6 = *applies fully*). Along with these three substantive trait scales, covariates were available in the datasets, namely, the score of an integrity test, the score of an achievement motivation test, and a measure of intelligence (aggregate of verbal, numeric, and figural cognitive ability test scores).

Pilot Studies

Before fitting models to the data, we ran a series of pilot studies to determine scoring weights for the faking dimension in the different personnel selection settings. Therefore, we asked independent samples of participants to rate the social desirability of every response category of every item with respect to the three job application settings (Seitz et al., 2025; see also Kuncel & Tellegen, 2009). Participants should hereby put themselves into the perspective of a person who is currently applying for the respective job, and rate desirability accordingly. The Supplemental Material provides details on the procedure, samples, and results of the pilot studies.

Results of the Empirical Demonstration

We fitted the same four models as in the simulation to all three datasets. Regarding scoring weights, we specified values as in Equation (2) and used the job-specific mean desirability ratings from the pilot studies as scoring weights of faking, which we linearly transformed to a possible range from 0 to 6 to achieve a common metric of scoring weights across dimensions. Again, we used JAGS through the R environment to estimate the models. However, for the empirical analyses, we estimated every model by running 12 parallel MCMC chains that featured 15,000 iterations each, with the first 5,000 iterations discarded as burnin.[3] $\hat{R}$ values of continuous model parameters were all below 1.1, indicating that all models converged. Also, by visual inspection, we found well-mixed trace plots of item parameters, latent correlations, and latent regression coefficients.

Model Selection and Model Fit. Table 4 contains fit indices of the four estimated models. The pattern of results was consistent across the three datasets. Regarding relative fit, DIC, WAIC, and LOOIC all selected the ''mixture θ_S/θ_F model,'' indicating that the mixture model in all datasets yielded a better compromise between fit and parsimony than the three non-mixture models. Regarding absolute fit, we used posterior predictive model checking (PPMC; e.g., Sinharay et al., 2006), which involves

Table 4. Empirical Demonstration: Model Fit Indices.

Model	Model selection criterion			SRMR
	DIC	WAIC	LOOIC	
Dataset 1 (bank apprenticeship)				
"θ_F model"	273,890.3	272,099.5	273,791.2	0.136
"θ_S model"	260,497.4	254,836.5	258,927.7	0.076
"θ_S/θ_F model"	256,858.1	249,733.1	254,442.7	0.066
"Mixture θ_S/θ_F model"	**254,427.1**	**246,881.8**	**251,235.1**	**0.056**
Dataset 2 (police officer traineeship)				
"θ_F model"	158,887.7	157,312.9	158,261.5	0.142
"θ_S model"	150,546.9	146,721.1	149,101.2	0.101
"θ_S/θ_F model"	146,888.4	142,359.9	144,863.4	0.064
"Mixture θ_S/θ_F model"	**145,447.3**	**140,678.5**	**142,999.4**	**0.059**
Dataset 3 (position as insurance agent)				
"θ_F model"	164,408.8	163,075.1	163,994.7	0.126
"θ_S model"	157,679.1	153,093.9	155,437.5	0.121
"θ_S/θ_F model"	150,547.5	145,811.4	148,618.3	0.072
"Mixture θ_S/θ_F model"	**146,579.0**	**141,219.6**	**143,669.9**	**0.061**

Note. Dataset 1: $N = 3,046$; Dataset 2: $N = 1,824$; Dataset 3: $N = 2,007$. The best-fitting model within each dataset is printed in bold. DIC = deviance information criterion; WAIC = widely applicable information criterion; LOOIC = leave-one-out information criterion; PPMC = posterior predictive model checking; SRMR = standardized root mean square residual (based on PPMC).

simulating data based on the model parameters' posterior distribution and comparing the simulated data to the observed data. As a measure of misfit, we considered the standardized root mean square residual (SRMR), which indicates the discrepancy between the model-implied and observed item intercorrelations.[4] The ''θ_F model'' had the largest SRMR in all job application contexts, followed by the ''θ_S model'' and ''θ_S/θ_F model'' (see Table 4). Crucially, the ''mixture θ_S/θ_F model'' consistently yielded the smallest SRMR, indicating that the mixture model had the best absolute fit in the three datasets.

Class Proportions. Looking at the estimated class proportions in the ''mixture θ_S/θ_F model,'' we found that every class had a considerable size in all datasets (see Table 5). Classes were, however, not equally sized. In all datasets, the ''S&F class'' had the largest class proportion (46.8–57.4%), the ''S-only class'' was the second largest class (27.0–44.0%), and the ''F-only class'' made up the smallest class (9.2–15.6%). Estimates of class proportions did not differ much between the three job application contexts.

Additionally, we examined classification diagnostics of the ''mixture θ_S/θ_F model,'' namely, posterior class probabilities and model entropy. These diagnostics give an indication about the (un)certainty of class assignments in mixture models (Masyn, 2013). In terms of posterior class probabilities, we considered for each test-

Table 5. Empirical Demonstration: Estimated Class Proportions and Model Entropies.

Dataset	Class			Entropy
	"S&F class"	"S-only class"	"F-only class"	
Dataset 1 (bank apprenticeship)	0.520 [0.468, 0.572]	0.384 [0.345, 0.423]	0.096 [0.077, 0.117]	0.854
Dataset 2 (police officer traineeship)	0.468 [0.407, 0.529]	0.440 [0.398, 0.482]	0.092 [0.068, 0.118]	0.892
Dataset 3 (position as insurance agent)	0.574 [0.514, 0.632]	0.270 [0.232, 0.310]	0.156 [0.125, 0.190]	0.900

Note. Dataset 1: $N = 3,046$; Dataset 2: $N = 1,824$; Dataset 3: $N = 2,007$. Values in brackets represent the 95% credible interval.

taker the percentage with which the modal class (i.e., the estimated class of a test-taker) was sampled during the MCMC estimation. The mean posterior class probabilities across test-takers were high for the three classes in all datasets (Dataset 1: 93.1–94.2%; Dataset 2: 94.7–95.5%; Dataset 3: 94.0–96.8%). Posterior class probabilities (π_{nc}) can be condensed in the measure of entropy, which has a range from 0 to 1 and is computed as $1 - \frac{\sum_{n=1}^{N}\sum_{c=1}^{3}(-\pi_{nc}\log(\pi_{nc}))}{N\log(3)}$ for the described mixture model. As noted in Table 5, entropies were also high in all datasets (0.854–0.900; cf. Vermunt, 2010).

Latent Regression of Class Membership. As mentioned above, the datasets also contained covariates. We included these variables as latent predictors of class membership in the "mixture θ_S/θ_F model" in all datasets. Table 6 shows the estimates of latent regression slopes. Results differed slightly between the datasets but were in general consistent. Higher integrity scores were associated with higher probabilities of being a member of the "S-only class" compared to being a member of the "S&F class" (i.e., the reference class). At the same time, higher integrity scores were generally associated with lower probabilities of being an "F-only class" member compared to being an "S&F class" member. For achievement motivation, results were inconsistent. Both among applicants for a bank apprenticeship and among applicants for a position as insurance agent, achievement motivation scores did not consistently predict class membership. Among applicants for a police officer traineeship, however, achievement motivation was negatively related to being an "S-only class" member and positively related to being an "F-only class" member. Similar to integrity, test-takers with higher scores of intelligence were more likely to belong to the "S-only class" compared to the "S&F class," whereas higher intelligence was associated with a lower probability of belonging to the "F-only class" compared to the "S&F class."

Table 6. Empirical Demonstration: Estimated Latent Regression Slopes.

	Dataset		
	Dataset 1 (bank apprenticeship)	Dataset 2 (police officer traineeship)	Dataset 3 (position as insurance agent)
“S-only class” vs. “S&F class” (reference class)			
Integrity	1.45 [1.21, 1.71]	0.75 [0.47, 1.05]	0.38 [0.13, 0.65]
Achievement motivation	0.29 [0.10, 0.49]	−1.16 [−1.43, −0.91]	0.08 [−0.29, 0.46]
Intelligence	1.31 [1.13, 1.48]	0.75 [0.56, 0.94]	0.59 [0.43, 0.75]
“F-only class” vs. “S&F class” (reference class)			
Integrity	−1.13 [−1.56, −0.71]	−0.65 [−1.07, −0.22]	−0.18 [−0.47, 0.12]
Achievement motivation	0.00 [−0.32, 0.34]	0.89 [0.59, 1.19]	0.00 [−0.42, 0.42]
Intelligence	−1.39 [−1.65, −1.14]	−0.59 [−0.85, −0.34]	−0.36 [−0.52. −0.20]

Note. Dataset 1: $N = 3,046$; Dataset 2: $N = 1,824$; Dataset 3: $N = 2,007$. All predictors were z-standardized. Positive slopes indicate that higher predictor values go along with higher probabilities of being a member of the “S-only class” or “F-only class” compared to the “S&F class.” Values in brackets represent the 95% credible interval.

Validation of Class Assignments

Response Distributions Within Classes. To investigate whether the classes indeed represented the response strategies outlined above, we analyzed response distributions within the three classes. For items at which the highest response category is most desirable, one can expect test-takers responding solely based on a faking dimension (i.e., the ‘‘F-only class’’) to yield higher mean responses than test-takers considering substantive traits and faking (i.e., the ‘‘S&F class’’). Test-takers only responding based on substantive traits (i.e., the ‘‘S-only class’’) should in turn yield the lowest mean responses. However, for items at which desirability does not increase monotonically with higher response categories, different effects can be expected. For items having their category of highest desirability above the midpoint though not at the extreme of the rating scale, differences in mean responses between the classes should be less pronounced compared to items at which the highest category is most desirable. In contrast, there should be no substantial mean differences between classes for items having their highest-desirability category at the scale midpoint. Figure 5 shows the class-specific distributions of mean item responses for the exemplary case of Dataset 1 (the pattern for the other datasets looked very similar). Results were generally in line with expectations, which supports the plausibility of class assignments in the ‘‘mixture θ_S/θ_F model.’’

Class Assignment of New Cases. To provide further evidence that the M-MNRM can afford a valid classification of a test-taker’s response strategy based on his or her response pattern, we applied the fitted ‘‘mixture θ_S/θ_F model’’ to another dataset that

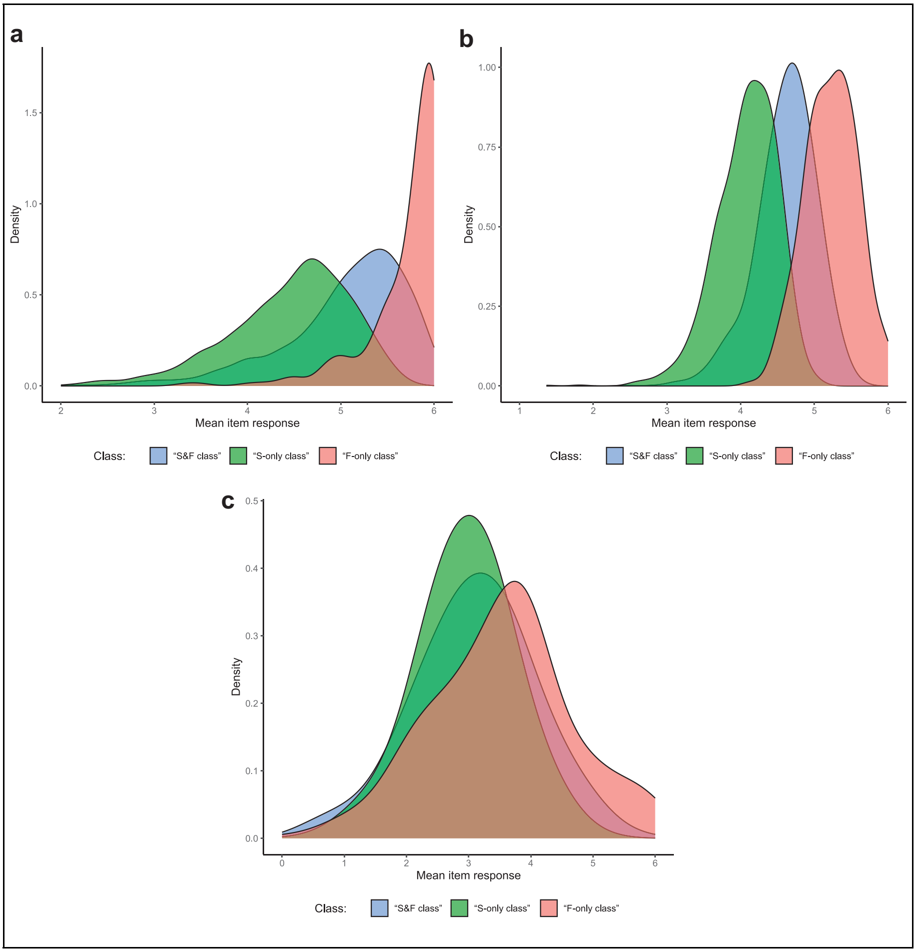


Figure 5. Empirical Demonstration: Class-Specific Distributions of Mean Item Responses for Items with Different Desirability Characteristics. (a) Items With a Highest-Desirability Category of "6"; (b) Items With a Highest-Desirability Category of "5"; (c) Items With a Highest-Desirability Category of "2," "3," or "4."

Note. Exemplary illustration for Dataset 1 (job application for a bank apprenticeship). The pattern for the other datasets was analogous. Plots display kernel densities of mean item responses, split by the three classes test-takers were assigned to by the "mixture θ_S/θ_F model" ("S&F class": $n = 1,597$, "S-only class": $n = 1,161$, "F-only class": $n = 288$). (a) Depicts the distributions for items at which the response category with highest desirability in the pilot study was "6," (b) for items with a highest-desirability category of "5," and (c) for items with a highest-desirability category of "2," "3," or "4" (see the Supplemental Material for more information on the pilot study).

was made available by the testing company. This dataset consisted of $N = 306$ subjects (gender: 49.7% female, 48.4% male; age: $M = 35.05$ years, $SD = 11.61$, range = [18, 64]) who responded to the same personality items used in our empirical analyses above. However, responses were not given in a hiring setting but in the context of career counseling. The personality test was embedded in a series of assessments based on which subjects received suggestions regarding suitable jobs and vocational paths. That is, one can expect that a considerably smaller proportion engages in faking in such a context, because honest responding is essential for effective career counseling and because responses are not used for selection purposes.

For this validation analysis, we fixed item parameters and latent correlations to the estimated values from the ''mixture θ_S/θ_F model'' in Dataset 3 (see Wetzel et al., 2021, for a similar approach),[5] and estimated the class proportions in the career counseling dataset. Because the covariates from the high-stakes datasets were not part of the career counseling assessment, we did not include any predictors in the latent regression of class membership. The estimated class proportions were 18.5% for the ''S&F class'' (95% credible interval: [14.1%, 23.4%]), 80.7% for the ''S-only class'' [75.8%, 85.1%], and 0.8% for the ''F-only class'' [0.1%, 2.1%]. That is, compared to the high-stakes context, the model assigned a much smaller proportion of subjects to classes including a faking dimension. The majority of subjects were instead classified as following a response strategy without faking, which is in line with expectations.

General Discussion

In this article, we proposed a mixture modeling extension of the MNRM to account for qualitatively different response strategies in high-stakes personality assessments. In the non-mixture MNRM, all test-takers are assumed to engage in some degree of faking, which influences item responses along with substantive traits (Seitz et al., 2024, 2025). Faking is represented as a quantitative difference variable in this model. In our mixture extension, however, faking is modeled in terms of both a continuous and discrete variable. This is in line with Kiefer and Benit's (2016) conclusion that a combined use of quantitative and qualitative modeling techniques would fit the current understanding of faking behavior best (cf. also Ziegler et al., 2015). In the M-MNRM, the discrete nature of faking is modeled by a latent class that represents conceivable response strategies in high-stakes assessments. The continuous nature of faking is modeled by a quantitative latent variable that represents the degree of aligning responses with desirability. Importantly, as opposed to the non-mixture MNRM, the M-MNRM is parameterized such that the quantitative faking variable only influences item responses if a test-taker adheres to a measurement model that includes a faking dimension.

Summary of Results

In the simulation study, we evaluated the M-MNRM in terms of parameter recovery compared to alternative non-mixture models. Overall, we found the M-MNRM to be superior to models without mixture components when there are indeed multiple classes in the data. Four points are worth emphasizing: First, parameter recovery in the M-MNRM was fairly stable across the different class proportion conditions. That is, the M-MNRM does not seem to require roughly equal class sizes but can outperform non-mixture models also if one class is much larger than the other classes or if one class is completely absent. Second, classification accuracy of the M-MNRM was high in all conditions, indicating that the model can correctly categorize response patterns as stemming from one of the described response strategies. Third, along with a better recovery of item parameters and latent correlations, the M-MNRM also improved the estimation of individual substantive trait scores, as the estimation of individual person parameters is based on a class-specific measurement model.[6] Fourth, covariate effects of different sizes were recovered well by the M-MNRM, such that substantive relationships between response strategy use and variables of interest can be modeled and tested effectively.[7]

At the same time, when there is only one class in the data, we found that the M-MNRM is not inferior to the non-mixture model representing the underlying population model. That is, although overparameterized, the M-MNRM does not seem to introduce bias or afford less precise estimates in non-mixture populations. This can be explained by the fact that the M-MNRM fitted to a non-mixture population assigns virtually all test-takers to the respective single class, such that the same set of data is used for parameter estimation in both the mixture and non-mixture model. Nevertheless, the M-MNRM constitutes an overly complex model when there is only one class in the data. Hence, for the sake of model parsimony and to reduce the risk of overfitting, a non-mixture model should be preferred in this case. Researchers and practitioners have two options to decide which model to choose in a given dataset: First, formal model selection criteria (DIC, WAIC, LOOIC) can be used. As opposed to model comparisons using likelihood-ratio tests in frequentist settings, DIC, WAIC, and LOOIC are no significance tests indicating whether a model fits the data significantly better than a more parsimonious model. Instead, they are information criteria that quantify the balance between mere model fit and model parsimony, which is achieved by penalizing model fit with the effective number of parameters or by considering out-of-sample predictive accuracy. As information criteria are descriptive measures, there are no fixed cutoffs or rules of thumb for differences in DIC, WAIC, or LOOIC to be considered meaningful. Instead, common practice is to simply select the model with the lowest information criterion value. In the simulation, when the population model was a mixture model, information criteria correctly chose the mixture model in all replications; when the population model was a non-mixture model, information criteria correctly chose a non-mixture model in more than 90% of the replications. Considering information criteria can hence be a trustworthy approach for choosing between models. Second, researchers and practitioners can consider the estimated class sizes in the M-MNRM for model selection. As indicated by the simulation, the M-MNRM can accurately estimate class proportions and test-takers' class membership under different kinds of data-generating models. Importantly, the simulation showed that this is also the case when one or two classes are truly absent in the data, as the proportion of empty classes is then indeed estimated to be 0 (see Alagöz & Meiser, 2024; Kim & Bolt, 2021; Tijmstra et al., 2018, who observed the same result for other mixture models estimated in a Bayesian framework). Given that DIC, WAIC, and LOOIC occasionally yielded false model selections in the simulation, we advise researchers and practitioners to pay close attention to how many test-takers actually make up each class, and let this information guide the decision of which model to choose: If test-takers are distributed across all classes, the full M-MNRM is appropriate; if one class is virtually empty, it is appropriate to not include the particular class; and if a single class contains virtually all test-takers, the respective non-mixture model is appropriate.

In the empirical demonstration, we showed that the M-MNRM can also prove successful in real high-stakes assessment data. We found the M-MNRM to be selected over non-mixture models in three datasets from different job application contexts, despite its higher complexity in terms of additional parameters. Also with regard to absolute fit, PPMC analyses revealed that the M-MNRM could describe the data better than non-mixture models. Furthermore, entropy values indicated good class separation in all empirical datasets. Comparing the results of the M-MNRM between the three job application contexts, it should be noted that there were no pronounced differences. Estimates of class proportions, for instance, were fairly constant in all datasets, suggesting that the non-mixture MNRM is misspecified for about 50% of test-takers in high-stakes assessments. For approximately 40%, a measurement model including only substantive traits seems to be more appropriate, whereas a one-dimensional measurement model of faking seems to describe the response behavior best for about 10%. Moreover, class membership was consistently predicted by integrity and intelligence, such that test-takers in the ''S-only class'' had the highest integrity and intelligence values and test-takers in the ''F-only class'' had the lowest values. Validation analyses also provided evidence for the plausibility of empirical class assignments performed by the M-MNRM, as class-specific response distributions and class proportions in a low-stakes sample were in line with expectations. To sum up, the consistent results across the three independent samples of job applicants provide evidence for the general usefulness of the M-MNRM in high-stakes personality assessments.

Utility of the Model

As mentioned in the introduction of the model, the M-MNRM constitutes a confirmatory as opposed to an exploratory mixture model. Jeon (2019) argued that, whereas exploratory mixture models allow researchers to *explore* the potential presence and substantive nature of multiple latent classes in the data, confirmatory mixture models are suited to *confirm* the existence and attributes of hypothesized latent classes. Similar to the distinction between exploratory and confirmatory factor models, the confirmatory approach in mixture modeling comes with stronger assumptions in the form of a-priori-constrained parameters. However, this precise definition of classes (a) allows for a theory-driven modeling of the data, (b) facilitates the interpretability of results, and (c) alleviates the risk of exploiting noise in the data (i.e., overfitting; cf. Celeux et al., 2019; van Havre et al., 2015). In our case, we modeled the three conceptual response strategies of how test-takers may or may not consider substantive traits and faking (Robie et al., 2007) and found all three classes to be present in empirical high-stakes data despite their restrictive definition with item slopes of non-used dimensions fixed to 0.

In applied settings, the combination of qualitative and quantitative modeling techniques allows practitioners (a) to make individual classifications regarding the response strategy of test-takers and (b) to more properly estimate and report substantive trait scores. In terms of individual response strategy classifications, the M-MNRM offers a faking detection technique that is more sophisticated than other indirect methods for detecting faking (e.g., LaHuis & Copeland, 2009; Sun et al., 2022; see Goldammer et al., 2024) or lie scales (e.g., Paulhus, 1988). Accurate faking classifications are vital for personality assessments used in high-stakes contexts. For instance, if test-takers have a very high probability of belonging to the ''F-only class,'' decision-makers should be informed that classical test scores, such as sum scores of raw item responses, are likely no valid indicators of the intended-to-be-measured traits. Likewise, accurate classifications are essential for the estimation of substantive trait scores in the two remaining classes to be based on the correct measurement model. As the simulation showed, the M-MNRM can indeed afford high hit rates and, consequently, improve the estimation of substantive trait scores compared to non-mixture models (more information on the class-specific estimation of person parameters can be found in the Supplemental Material).

Limitations and Future Research Directions

Some limitations as well as future research directions warrant mentioning. One caveat concerns the direct comparability of person parameter estimates across classes, which can be an issue in all types of latent class models. To be able to make meaningful comparisons of person parameters across classes, parameters must be on a common scale (Paek & Cho, 2015). In the M-MNRM as presented in this article, item slopes of dimensions not fixed to 0 are constrained to be class-invariant, in order to allow for the same latent variables to be measured in every class. Item-category intercepts, however, are unconstrained between classes, in order to capture different response distributions in the three classes, which should facilitate class separation when estimating the model. There are different approaches for establishing a common scale of person parameters across classes despite this non-invariance of intercepts. These resemble scaling methods in the context of test equating (see Kolen & Brennan,

2004) as well as approaches for creating a common scale of item difficulty parameters in mixture Rasch models (Paek & Cho, 2015; Rost, 1990). One option is to model a set of so-called anchor items. Anchor items are items whose slope and intercept parameters are the same in all classes, such that these items create a common scale of person parameters. In the context of faking, items with neutral social desirability could serve as viable anchor items. Alternatively, assuming that the response strategies test-takers employ in a high-stakes assessment do not transfer to an assessment context where stakes are low, items administered to test-takers in a low-stakes assessment setting could be candidate anchor items. If no anchor items are available (as in our empirical demonstration), a common scale of person parameters can also be achieved if the latent classes follow the same true trait distribution. For model identification, latent means and variances of the multivariate normal distribution of person parameters are set to 0 and 1, respectively, for all classes in the M-MNRM. If this class-invariant identification of the scale of person parameters is in line with the truth (i.e., if classes truly do not differ in their trait distribution), person parameter estimates should be on the same scale across classes and hence be comparable. Such a situation was modeled in an additional simulation reported in the Supplemental Material, where intercepts were systematically different between classes but person parameters were drawn from the same multivariate normal distribution irrespective of class membership. As illustrated in Figure S2, non-mixture models in such a scenario yield pronounced mean differences between classes in substantive trait score estimates. In contrast, the M-MNRM does not produce such a bias, allowing that test-takers can be meaningfully ranked across classes on substantive traits.[8] The assumption of a homogeneous trait distribution across classes can be appropriate for many personality constructs. However, if this assumption is violated, a common scale of person parameters will not be achieved unless the scale is matched by other test equating methods (such as anchor items; see Kolen & Brennan, 2004). Hence, we generally advise researchers and practitioners to compare person parameters of test-takers from different classes with caution.

Another set of limitations is related to the simulation of the current article, which featured different class proportion conditions but was limited to a fixed sample size and test length. Even though our simulation design was representative of datasets in empirical high-stakes settings (see the empirical demonstration above), future research should study the performance of the M-MNRM in data situations with different numbers of test-takers, items, and substantive trait scales. Also, even though monotonically increasing, nonmonotonically increasing, as well as inverted-U-shaped relations between response categories and desirability were simulated, scoring weights of faking were equidistant within the segments of the desirability trajectories in our simulation (see Figure 1 in Seitz et al., 2024). In empirical settings, however, the relation between categories and desirability may well take on idiosyncratic forms (Kuncel & Tellegen, 2009). To assess the sensitivity of the M-MNRM to situations where categories are related to desirability in idiosyncratic, non-equidistant ways, we ran an additional simulation in which we set scoring weights of faking to desirability values collected in the pilot study for Dataset 1 of the empirical demonstration (see the Supplemental Material for details). This should emulate realistic relations between categories and desirability, where scoring weights of faking are not necessarily equidistant. Results of this additional simulation were essentially the same as in the main simulation with equidistant scoring weights of faking. Nevertheless, future research could examine the M-MNRM's sensitivity to different kinds of item desirability characteristics in more detail (cf. Seitz et al., 2024, who conducted a similar investigation in the context of the non-mixture MNRM). Additionally, the number of replications per condition in our simulation was limited to 30. This was primarily due to the computational complexity of the M-MNRM (the estimation of the model in a single replication of the simulation took more than ten hours on a high-performance computer). As indicated by the error bars in Figures 1 to 4, results were nevertheless fairly reliable with 30 replications per condition. The model's computational complexity with its Bayesian estimation may in itself constitute a limiting factor for applied researchers and practitioners who wish to apply the model. However, this limitation will be alleviated in the future with the ever-increasing availability of high-performance computing machines. To facilitate dissemination of the model, we provide commented syntax files for estimating the M-MNRM in an exemplary dataset (available at https://osf.io/vwqf3/).

Also, notwithstanding the above-described positive features of confirmatory mixture models, we encourage future research to model different response strategies in high-stakes assessments in a less restrictive manner, for instance, by specifying class- and dimension-specific proportionality constants on a class-invariant matrix of item slopes. Future studies could also test alternative parameterizations of the ''F-only class,'' for example, one with a model of independence. In such a model, ''F-only'' test-takers' overall tendency to respond according to desirability characteristics would only be captured by item-category intercepts, whereas no variance would be explained by a common factor in this class (i.e., all variation would be unsystematic). However, it might be difficult to separate such a class from a class of test-takers who simply respond inattentively (cf. Jin et al., 2018), even though careless responding is arguably very rare in high-stakes assessments. An additional extension of the model would be to allow class membership to vary not only between persons but also between items. This would further increase the flexibility of the model and account for switches between response strategies over the course of the questionnaire. However, it should be noted that the information for class assignments in such a person-by-item mixture model would be very sparse compared to a person mixture model (namely, single item responses instead of a whole response vector). To overcome this challenge, it could be worthwhile incorporating external information, such as response times or other process data (see Ulitzsch et al., 2022, who modeled careless responding with a person-by-item mixture model that included response times). In such a model, item-level covariates like item wording or other item characteristics could be modeled to examine which types of item content are particularly susceptible to faking.

Furthermore, it would be appealing to compare the faking detection accuracy of the M-MNRM to the accuracy of other recent faking detection methods, such as machine-learning-based approaches (Calanna et al., 2020; see also Nie et al., 2025) or approaches using IRTree models (Sun et al., 2022). Seeing under which conditions the different methods perform best could help develop an integrative faking detection technique that combines features from mixture IRT(ree) modeling and machine learning. Another interesting endeavor for future studies analyzing follow-up data from hired applicants would be to link test-takers' class membership in the M-MNRM to real-world job performance outcomes, especially to contextual performance and counterproductive work behavior. In the empirical demonstration of the current article, class membership was associated with integrity, achievement motivation, and intelligence. If, however, membership in the ''S-only class'' was related to actual organizational citizenship behavior, or if membership in the ''F-only class'' predicted undesired actions like turnover or absenteeism, distinguishing between different response strategies in high-stakes assessments would not only be important from a psychometric measurement perspective but also provide in itself a valuable piece of information for hiring decisions. Investigating associations between class membership and meaningful consequences on the job would hence be helpful to showcase the utility of such mixture models in applied measurement contexts like personnel selection.

Conclusion

To conclude, as our simulation and empirical demonstration illustrated, the M-MNRM provides a valuable extension of latent variable models of faking. Compared to many other faking models, the M-MNRM is not restricted to a single measurement model but allows for qualitatively different response strategies employed by test-takers. Both psychometrically and from an applied measurement perspective, such an extension is worthwhile since it offers researchers and practitioners a tool to detect and account for response strategies associated with a different use of substantive traits and faking, which would otherwise bias results. Future research can help to discover boundary conditions of the model's efficacy or alternative parameterizations that allow a sophisticated modeling of different faking tendencies.

Acknowledgments

The authors cordially thank HR Diagnostics for providing the data for the empirical demonstration of this article.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—GRK 2277—Projektnummer 310365261.

Ethical Considerations

Not applicable because the study involves a statistical simulation and a reanalysis of existing datasets.

Consent to Participate and for Publication

Not applicable because the study involves a statistical simulation and a reanalysis of existing datasets.

ORCID iD

Timo Seitz https://orcid.org/0000-0002-7375-4511

Availability of Data and Materials

The JAGS model syntaxes, exemplary *R* code for estimating the M-MNRM, as well as the instructions, a screenshot, the data, and the analysis script of the pilot studies are available at https://osf.io/vwqf3/

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Because class membership was based on the described sampling procedure, slight deviations between the actual class proportions in a replication and the class proportions presented in the Simulation Conditions section could occur. However, deviations were very small ($M = 0.000$, $SD = 0.014$, range $= [-0.060, 0.059]$). Also, it was made sure that classes that should be absent by design were indeed empty by using regression intercepts of $-\infty$.
2. Due to data protection guidelines, there was no age variable in Dataset 3. According to consultations with the testing company, the approximate average age of applicants for this particular position was above 30 years, with a range from 20 to 45.
3. As is well known in the psychometric literature, mixture models can be hard to estimate in empirical data because of multimodal likelihood functions (e.g., Hipp & Bauer, 2006; McLachlan & Peel, 2000). Instead of just discarding MCMC chains that converged to a local solution, we employed a two-step estimation of the ''mixture θ_S/θ_F model.'' The first step should optimize initial values for the actual model estimation in the second step (cf. O'Hagan et al., 2012). This was done by running 30 parallel chains in the first step (each

with 2,000 iterations after a 2,000-iteration burnin phase) and using the parameter estimates from the chain that yielded the highest model likelihood as initial values for the second step. The second step then featured the described estimation with 12 parallel chains. However, to prevent that all chains started with the exact same set of initial values, we added random noise to the initial values of every chain. Details can be found at https://osf.io/vwqf3/.

4. Specifically, for each MCMC iteration after the burnin phase, item responses were simulated using the model equation of the respective model and the parameters sampled in the given iteration. Then, the root of the mean squared deviation between item intercorrelations in the simulated versus observed data was computed for each iteration, before the average across iterations was calculated to yield the SRMR of a model.
5. We chose Dataset 3 for the validation analysis because item wordings in the career counseling dataset were not fully identical to item wordings in Datasets 1 and 2. Some item wordings in Datasets 1 and 2 had been adjusted to better fit the age group of applicants for the jobs (mainly high school graduates).
6. More information on the class-specific estimation of person parameters can be found in the Supplemental Material. There, we also report the debiasing effect of the mixture model on substantive trait scores in a simulation with item-category intercepts differing systematically between classes.
7. We also conducted a simulation with all covariate effects fixed to 0 in the data generation (see the Supplemental Material). Results, including the recovery of class proportions, class membership, and latent regression coefficients, were equivalent to the findings reported above. This suggests that the M-MNRM does not require class membership predictors of considerable effect size to produce satisfactory results.
8. Note that it only makes sense to interpret and rank substantive trait score estimates of test-takers from classes where item responses are influenced by substantive traits (i.e., ‘‘S&F class’’ and ‘‘S-only class’’) As described in the Supplemental Material, the M-MNRM formally estimates substantive trait scores for test-takers in the ‘‘F-only class’’ (which should be close to 0), but these estimates are meaningless and should not be interpreted.

References

Alagöz, Ö. E. C., & Meiser, T. (2024). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement*, *84*(5), 957–993. https://doi.org/10.1177/00131644231206765

Arthur, W. Jr., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality. *International Journal of Selection and Assessment*, *18*(1), 1–16. https://doi.org/10.1111/j.1468-2389.2010.00476.x

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. https://doi.org/10.1007/bf02291411

Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. *Psychometrika*, *79*(3), 515–537. https://doi.org/10.1007/s11336-013-9390-9

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

Brown, A., & Böckenholt, U. (2022). Intermittent faking of personality profiles in high-stakes assessments: A grade of membership analysis. *Psychological Methods*, *27*(5), 895–916. https://doi.org/10.1037/met0000295

Calanna, P., Lauriola, M., Saggino, A., Tommasi, M., & Furlan, S. (2020). Using a supervised machine learning algorithm for detecting faking good in a personality self-report. *International Journal of Selection and Assessment*, *28*(2), 176–185. https://doi.org/10.1111/ijsa.12279

Celeux, G., Frühwirth-Schnatter, S., & Robert, C. P. (2019). Model selection for mixture models—Perspectives and strategies. In *Handbook of mixture analysis* (pp. 117–154). Chapman & Hall/CRC.

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

Christiansen, N. D., Robie, C., Burns, G. N., Loy, R. W., Speer, A. B., & Jacobs, R. R. (2021). Effects of applicant response distortion on the relationship between personality trait scores and cognitive ability. *Personality and Individual Differences*, *171*, Article 110542. https://doi.org/10.1016/j.paid.2020.110542

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*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics*, *23*(4), 643–659. https://doi.org/10.1007/s00180-007-0103-7

Diekmann, J., & König, C. J. (2015). Personality testing in personnel selection: Love it? Leave it? Understand it! In I. Nikolaou & J. K. Oostrom (Eds.), *Employee recruitment, selection, and assessment* (pp. 129–147). Psychology Press.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, *16*(1), 81–106. https://doi.org/10.1207/S15327043HUP1601_4

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment*, *16*(1), 20–30. https://doi.org/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*(3), 328–347. https://doi.org/10.1037/met0000059

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall/CRC.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Goldammer, P., Stöckli, P. L., Escher, Y. A., Annen, H., & Jonas, K. (2024). On the utility of indirect methods for detecting faking. *Educational and Psychological Measurement*, *84*(5), 841–868. https://doi.org/10.1177/00131644231209520

Gollwitzer, M., Eid, M., & Jürgensen, R. (2005). Response styles in the assessment of anger expression. *Psychological Assessment*, *17*(1), 56–69. https://doi.org/10.1037/1040-3590.17.1.56

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review*, *36*(3), 341–355. https://doi.org/10.1108/00483480710731310

Griffith, R. L., & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0018

Hendy, N., Krammer, G., Schermer, J. A., & Biderman, M. D. (2021). Using bifactor models to identify faking on Big Five questionnaires. *International Journal of Selection and Assessment*, *29*(1), 81–99. https://doi.org/10.1111/ijsa.12316

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

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, *11*(1), 36–53. https://doi.org/10.1037/1082-989X.11.1.36

Jeon, M. (2019). A specialized confirmatory mixture IRT modeling approach for multidimensional tests. *Psychological Test and Assessment Modeling*, *61*(1), 91–123.

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Mixture item response models for inattentive responding behavior. *Organizational Research Methods*, *21*(1), 197–225. https://doi.org/10.1177/1094428117725792

Kanfer, R., Wanberg, C. R., & Kantrowitz, T. M. (2001). Job search and employment: A personality–motivational analysis and meta-analytic review. *Journal of Applied Psychology*, *86*(5), 837–855. https://doi.org/10.1037/0021-9010.86.5.837

Kiefer, C., & Benit, N. (2016). What is applicant faking behavior? A review on the current state of theory and modeling techniques. *Journal of European Psychology Students*, *7*(1), 9–19. https://doi.org/10.5334/jeps.345

Kim, N., & Bolt, D. M. (2021). A mixture IRTree model for extreme response style: Accounting for response process uncertainty. *Educational and Psychological Measurement*, *81*(1), 131–154. https://doi.org/10.1177/0013164420913915

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance*, *25*(4), 273–302. https://doi.org/10.1080/08959285.2012.703733

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). Springer.

König, C. J., Hafsteinsson, L. G., Jansen, A., & Stadelmann, E. H. (2011). Applicants' self-presentational behavior across cultures: Less self-presentation in Switzerland and Iceland than in the United States. *International Journal of Selection and Assessment*, *19*(4), 331–339. https://doi.org/10.1111/j.1468-2389.2011.00562.x

Kuncel, N. R., Borneman, M., & Kiger, T. (2011). Innovative item response process and Bayesian faking detection methods. In M. Ziegler, C. MacCann, & R. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 102–112). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0036

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x

LaHuis, D. M., & Copeland, D. (2009). Investigating faking using a multilevel logistic regression approach to measuring person fit. *Organizational Research Methods*, *12*(2), 296–319. https://doi.org/10.1177/1094428107302903

Leng, C. H., Huang, H. Y., & Yao, G. (2020). A social desirability item response theory model: Retrieve-deceive-transfer. *Psychometrika*, *85*(1), 56–74. https://doi.org/10.1007/s11336-019-09689-y

Marcus, B. (2009). ''Faking'' from the applicant's perspective: A theory of self-presentation in personnel selection settings. *International Journal of Selection and Assessment*, *17*(4), 417–430. https://doi.org/10.1111/j.1468-2389.2009.00483.x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/bf02296272

Masyn, K. (2013). Latent class analysis and finite mixture modeling. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (*Vol. 2*, pp. 551–611). Oxford University Press.

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, *88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

Nie, W., Hernandez, I., Tay, L., Zhang, B., & Cao, M. (2025). A comparison of the response-pattern-based faking detection methods. *Journal of Applied Psychology*. Advance online publication. https://doi.org/10.1037/apl0001261

Nikolaou, I., & Foti, K. (2018). Personnel selection and personality. In V. Zeigler-Hill & T. Shackelford (Eds.), *The SAGE handbook of personality and individual differences; Volume III: Applications of personality and individual differences* (pp. 458–474). SAGE Publications. https://doi.org/10.4135/9781526451248.n20

O'Hagan, A., Murphy, T. B., & Gormley, I. C. (2012). Computational aspects of fitting mixture models via the expectation—Maximization algorithm. *Computational Statistics & Data Analysis*, *56*(12), 3843–3864. https://doi.org/10.1016/j.csda.2012.05.011

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*(4), 995–1027. https://doi.org/10.1111/j.1744-6570.2007.00099.x

Paek, I., & Cho, S.-J. (2015). A note on parameter estimate comparability: Across latent classes in mixture IRT modeling. *Applied Psychological Measurement*, *39*(2), 135–143. https://doi.org/10.1177/0146621614549651

Paulhus, D. L. (1988). *Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding* [Unpublished manual]. Department of Psychology, University of British Colombia.

Plummer, M. (2017). *JAGS: Just another Gibbs sampler* (version 4.3.1) [Computer software]. https://sourceforge.net/projects/mcmc-jags/

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11.

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology*, *21*(4), 489–509. https://doi.org/10.1007/s10869-007-9038-9

Röhner, J., Schütz, A., & Ziegler, M. (2025). Faking in self-report personality scales: A qualitative analysis and taxonomy of the behaviors that constitute faking strategies. *International Journal of Selection and Assessment*, *33*(1), e12513. https://doi.org/10.1111/ijsa.12513

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271–282. https://doi.org/10.1177/014662169001400305

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology*, *44*(1), 75–92. https://doi.org/10.1111/j.2044-8317.1991.tb00951.x

Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science*, *9*(5), 538–551. https://doi.org/10.1177/1745691614543972

Saks, A. M., & Ashforth, B. E. (2002). Is job search related to employment quality? It all depends on the fit. *Journal of Applied Psychology*, *87*(4), 646–654. https://doi.org/10.1037/0021-9010.87.4.646

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, *78*(6), 966–974. https://doi.org/10.1037/0021-9010.78.6.966

Seitz, T., Spengler, M., & Meiser, T. (2025). ‘‘What if applicants fake their responses?’’: Modeling faking and response styles in high-stakes assessments using the multidimensional nominal response model. *Educational and Psychological Measurement*. Advance online publication. https://doi.org/10.1177/00131644241307560

Seitz, T., Wetzel, E., Hilbig, B. E., & Meiser, T. (2024). Using the multidimensional nominal response model to model faking in questionnaire data: The importance of item desirability characteristics. *Behavior Research Methods*, *56*(8), 8869–8896. https://doi.org/10.3758/s13428-024-02509-x

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*(4), 298–321. https://doi.org/10.1177/0146621605285517

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*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *62*(4), 795–809. https://doi.org/10.1111/1467-9868.00265

Sun, T., Zhang, B., Cao, M., & Drasgow, F. (2022). Faking detection improved: Adopting a Likert item response process tree model. *Organizational Research Methods*, *25*(3), 490–512. https://doi.org/10.1177/10944281211002904

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. https://doi.org/10.1007/bf02294363

Thissen, D., & Cai, L. (2016). Nominal categories models. In W. J. van der Linden (Ed.), *Handbook of item response theory, Volume 1: Models* (pp. 51–73). Chapman & Hall/CRC Press.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577. https://doi.org/10.1007/bf02295596

Tijmstra, J., Bolsinova, M., & Jeon, M. (2018). General mixture item response models with different item response structures: Exposition with an application to Likert scales. *Behavior Research Methods*, *50*(6), 2325–2344. https://doi.org/10.3758/s13428-017-0997-0

Turner, E. (2022, September 19). *Detection apprehension: The fear of being caught lying*. Paul Ekman Group. https://www.paulekman.com/blog/detection-apprehension/

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

van Havre, Z., White, N., Rousseau, J., & Mengersen, K. (2015). Overfitting Bayesian mixture models with an unknown number of components. *PLoS One*, *10*(7), e0131739. https://doi.org/10.1371/journal.pone.0131739

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, *18*(4), 450–469. https://doi.org/10.1093/pan/mpq025

von Davier, M., & Rost, J. (2006). Mixture distribution item response models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (*Vol. 26*, pp. 643–661). Elsevier. https://doi.org/10.1016/S0169-7161(06)26019-X

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial-credit model. *Applied Psychological Measurement*, *28*(6), 389–406. https://doi.org/10.1177/0146621604268734

von Davier, M., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 99–115). Springer. https://doi.org/10.1007/978-0-387-49839-3_6

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Wetzel, E., Frick, S., & Brown, A. (2021). Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking. *Psychological Assessment*, *33*(2), 156–170. https://doi.org/10.1037/pas0000971

Yamamoto, K. (1987). *A model that combines IRT and latent class models* [Unpublished doctoral dissertation]. University of Illinois at Urbana-Champaign.

Yamamoto, K. (1989). *Hybrid model of IRT and latent class models* (ETS Report RR-89-41). Educational Testing Service (ETS).

Yamamoto, K., & Everson, H. T. (1995). *Modeling the mixture of IRT and pattern responses by a modified hybrid model* (ETS Report RR-95-16). Educational Testing Service (ETS).

Youngflesh, C. (2018). MCMCvis: Tools to visualize, manipulate, and summarize MCMC output. *Journal of Open Source Software*, *3*(24), 640. https://doi.org/10.21105/joss.00640

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, *7*(2), 168–190. https://doi.org/10.1177/1094428104263674

Ziegler, M. (2011). Applicant faking: A look into the black box. *The Industrial and Organizational Psychologist*, *49*(1), 29–36.

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods*, *18*(4), 679–703. https://doi.org/10.1177/1094428115574518

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). Faking: Knowns, unknowns, and points of contention. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 3–16). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0011

Faking in High-Stakes Personality Assessments: A Response-Time-Based Latent Response Mixture Modeling Approach

Timo Seitz[1]; Esther Ulitzsch[2]

[1] University of Mannheim, Mannheim, Germany; [2] Centre for Educational Measurement (CEMO), University of Oslo, Oslo, Norway

Author Note

Timo Seitz https://orcid.org/0000-0002-7375-4511

Esther Ulitzsch https://orcid.org/0000-0002-9267-8542

The authors cordially thank HR Diagnostics for providing the data for the empirical demonstration of this article.

Correspondence concerning this manuscript should be addressed to Timo Seitz, Department of Psychology, University of Mannheim, L13,15-17 – room 515, 68161 Mannheim, Germany. E-mail: timo.seitz@uni-mannheim.de

(Statements and Declarations on next page)

Statements and Declarations

Funding Statement

This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – GRK 2277 – Projektnummer 310365261, and partially supported by the Research Council of Norway through its Centres of Excellence scheme, project number 33160.

Material and Data Availability

The JAGS model syntaxes, R code for estimating the proposed model, as well as the materials, data, and analysis scripts of the pre-study as well as pilot study are available at https://osf.io/crmv4/?view_only=db3eb1f7139742f48ee01887363b5e4a.

Declaration of Conflicting Interest

The authors have no potential conflicts of interest to declare concerning the research, authorship, and/or publication of this article.

Ethical Considerations

Not applicable because the study involves analyses of simulated datasets as well as a reanalysis of an existing empirical dataset

Consent to Participate and for Publication

Not applicable because the study involves analyses of simulated datasets as well as a reanalysis of an existing empirical dataset

Abstract

When personality assessments are employed in high-stakes contexts, there is the risk that test-takers provide overly positive descriptions of themselves. This response bias is known as faking and has often been addressed in latent variable models through an additional dimension capturing each test-taker's faking degree. Such models typically assume a homogeneous response strategy for all test-takers, with substantive traits and faking jointly influencing responses to all items. In this article, we present a latent response mixture item response theory (IRT) model of faking that accounts for changes in test-takers' response strategies over the course of the assessment. The model translates theoretical considerations about test-taker behavior into different model components for item responses and corresponding item-level response times (RT), thereby allowing to account for, identify, and investigate different faking-related response strategies on the person-by-item level. In a simulation study, we found that the model parameters can be estimated well under realistic conditions. Also, we applied the model to an empirical dataset ($N = 1824$) from a job application context, showcasing its utility in real high-stakes assessment data. We conclude the article by discussing the role of the model for psychological measurement as well as substantive research.

Keywords: faking, mixture model, response times, item response theory

Introduction

Personality questionnaires based on self-report are commonly used in high-stakes contexts like personnel selection or college admission, as personality traits have repeatedly been shown to predict relevant outcomes such as job performance or academic success (e.g., Sackett & Walmsley, 2014; Poropat, 2009). Nevertheless, employing self-report personality tests in contexts where the assessment results have important consequences for test-takers holds the risk that test-takers portray themselves in an overly positive light. This behavior is commonly referred to as *faking*, and its effects on the psychometric properties of a test are well documented (Ziegler et al., 2011). For instance, faking leads to negatively (positively) skewed distributions of items and scales measuring desirable (undesirable) attributes (e.g., Hu & Connelly, 2021). Also, it alters rank orders of test-takers (e.g., Mueller-Hanson et al., 2003), which impacts decisions where test-takers are selected based on their test scores. Most importantly, however, faking distorts construct validity of personality measures as it introduces an additional source of systematic variance. This leads to inflated inter-item and inter-scale correlations, diminishing the test's factor structure (e.g., Schmit & Ryan, 1993) and rendering the obtained scores inappropriate for their intended purpose of usage (Messick, 1989). Hence, it is vital to have psychometric tools that effectively account for faking, especially in the context of high-stakes assessments. Such methods can be used to detect and statistically control for faking, but also to study the associated response process. Whereas the former can pay dividends when data have already been collected, the latter is important for advancing the theoretical understanding of faking, which can in turn be used to develop assessment tools that are less susceptible to faking.

Faking is especially problematic because of its heterogeneity. This is evidenced by studies investigating the prevalence of faking in job applications. Griffith and Converse (2011) reviewed the literature on this topic and concluded that roughly 30% (±10%) of applicants engage in faking behavior. That is, some test-takers indeed respond in a way that facilitates their appearance, whereas others do not. This is supported by Robie et al. (2007) who conducted a think-aloud study of test-takers responding to a personality test in a high-stakes condition. Using a verbal protocol analysis, they found one group of test-takers being fully honest in their responses, one group considering both their actual personality and the criteria of an “ideal” applicant, as well as one group exclusively responding based on “ideal” applicant considerations. In a related study, Röhner et al. (2025) identified as many as 35 different behaviors constituting distinct faking strategies.

To account for heterogeneity in faking, several model-based approaches have been proposed. Most of them either quantify the degree of faking on a latent continuum (e.g., Klehe et al., 2012; Hendy et al., 2021; Seitz et al., 2024; Seitz, Spengler, & Meiser, 2025) or assign test-takers to latent classes that represent qualitatively different response behaviors related to faking (e.g., Zickar et al., 2004). Other models combine quantitative and qualitative conceptualizations of faking to better describe the nature of the construct (Seitz, Alagöz, & Meiser, 2025; Ziegler et al., 2015). What most models have in common, however, is that they treat faking as a person variable that is constant across test items (see Böckenholt, 2014; Brown & Böckenholt, 2022; for exceptions). Yet, one can for multiple reasons question whether it is appropriate to conceptualize faking as constant throughout the entire test: First, faking is a complex interaction of person and situation characteristics, such as ability, opportunity, and motivation to fake (Tett & Simonet, 2011). Motivation to fake in personnel selection, for instance, may vary between items when some items are perceived as more instrumental than others to elevate the chance of being hired in a given job context (e.g., Ellingson & McFarland, 2011). Second, lying research has shown that people tend to behave dishonestly to the extent that they profit, but not to the extent that they damage their self-concept of being an honorable person (Mazar et al., 2008). If one has nothing to conceal at a particular item such that misreporting would not pay off, it can hence be expected that people will not engage in faking (Tourangeau & Yan, 2007). Third, test-takers have conflicting goals in high-stakes assessments. On the one hand, they want to impress a prospective employer. On the other hand, conveying a credible impression and staying true to oneself have also been identified as important motives of test-taking behavior (Kuncel et al., 2011). All these arguments suggest it is more plausible to assume that faking also varies between items than to treat it as a constant person variable.

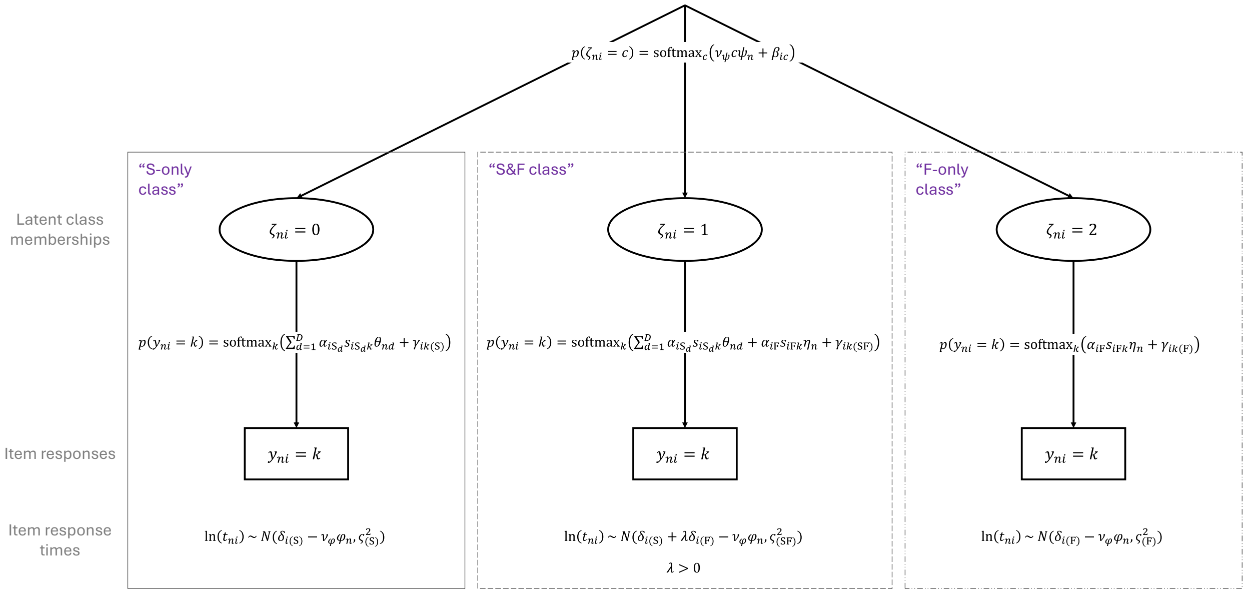
In this work, we present a mixture item response theory (IRT) model that allows to account for, identify, and investigate different faking-related response strategies on the person-by-item level. The model we propose translates theoretical considerations about test-taker behavior into different model components (i.e., latent classes) for item responses and corresponding item-level response times (RT). Modeling RTs serves two purposes: First, from a statistical perspective, incorporating additional behavioral data into the model facilitates class separation. Second, from a substantive perspective, doing so allows investigating the response process behind faking in a sophisticated manner. It, for instance, allows examining the question of whether faking actually increases or decreases RTs. Also, the model as a whole can be used to identify what items are especially susceptible to faking, or to study relationships of substantive person characteristics with different faking tendencies. The inclusion of RTs thus constitutes an important extension over existing models allowing for varying faking behavior across items. Furthermore, our proposed model has the advantage of being able to account for nonmonotonic faking effects (see Figure 2) as well as including an additional faking class (a more detailed description of differences to related models can be found below).

Proposed Model

The model we present builds on work by Seitz, Alagöz, and Meiser (2025), who modeled faking using three latent classes related to distinct response strategies test-takers can employ in high-stakes assessments. We extend their model by adding an item-wise mixture component and combine it with approaches that make use of RT data to account for disengaged and careless responding in the context of cognitive and non-cognitive assessments (e.g., Ulitzsch et al., 2020; Ulitzsch, Pohl, et al., 2022). The full model is displayed in Figure 1.

Figure 1

The Proposed Response-Time-Based Latent Response Mixture Model of Faking


$p(\zeta_{ni} = c) = \text{softmax}_c(\nu_\psi c\psi_n + \beta_{ic})$
Latent class memberships
Item responses
Item response times
"S-only class"
$\zeta_{ni} = 0$
$p(y_{ni} = k) = \text{softmax}_k(\sum_{d=1}^{D} \alpha_{i\text{S}_d} s_{i\text{S}_d k}\theta_{nd} + \gamma_{ik(\text{S})})$
$y_{ni} = k$
$\ln(t_{ni}) \sim N(\delta_{i(\text{S})} - \nu_\varphi\varphi_n, \varsigma^2_{(\text{S})})$
"S&F class"
$\zeta_{ni} = 1$
$p(y_{ni} = k) = \text{softmax}_k(\sum_{d=1}^{D} \alpha_{i\text{S}_d} s_{i\text{S}_d k}\theta_{nd} + \alpha_{i\text{F}} s_{i\text{F}k}\eta_n + \gamma_{ik(\text{SF})})$
$y_{ni} = k$
$\ln(t_{ni}) \sim N(\delta_{i(\text{S})} + \lambda\delta_{i(\text{F})} - \nu_\varphi\varphi_n, \varsigma^2_{(\text{SF})})$
$\lambda > 0$
"F-only class"
$\zeta_{ni} = 2$
$p(y_{ni} = k) = \text{softmax}_k(\alpha_{i\text{F}} s_{i\text{F}k}\eta_n + \gamma_{ik(\text{F})})$
$y_{ni} = k$
$\ln(t_{ni}) \sim N(\delta_{i(\text{F})} - \nu_\varphi\varphi_n, \varsigma^2_{(\text{F})})$

Model Components

Item Response Models

Similar to Seitz, Alagöz, and Meiser (2025), we assume that item responses are mixtures of three response strategies (see also the study by Robie et al., 2007, described above). The first response strategy ("S-only class") is a strategy where test-takers respond according to their substantive traits, that is, respond honestly and do not align responses with social desirability. The second response strategy ("S&F class") is a strategy where item responses are a function of test-takers' substantive traits as well as a faking dimension. Responses thus represent edited versions of honest responses depending on the faking degree of a test-taker. The third response strategy ("F-only class") is a strategy where item responses are independent of test-takers' substantive traits. Instead, responses are solely influenced by the faking dimension.

Item responses $y_{ni} \in \{0, 1, \dots, k, \dots, K\}$ are modeled using different IRT models. Technically, the different models are constrained versions of Falk and Cai's (2016) parameterization of the multidimensional nominal response model (MNRM; Takane & de Leeuw, 1987) depending on response strategy use. Strategy use per item is represented as a latent class $\zeta_{ni} \in \{0, 1, 2\}$. In the "S&F class"

($\zeta_{ni} = 1$), where responses are a function of substantive traits and faking, the probability of person n choosing response category k on item i is modeled as

$$p(y_{ni} = k \mid \zeta_{ni} = 1) = \text{softmax}_k\left(\sum_{d=1}^{D} \alpha_{i\mathrm{S}_d} s_{i\mathrm{S}_d k} \theta_{nd} + \alpha_{i\mathrm{F}} s_{i\mathrm{F}k} \eta_n + \gamma_{ik(\mathrm{SF})}\right). \quad (1)$$

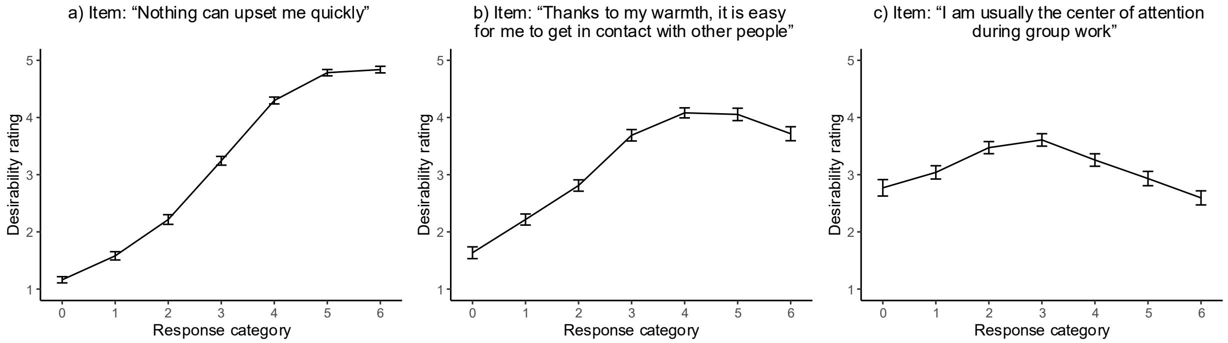
This softmax function (aka multinomial logistic function) converts a vector of $K + 1$ real-valued category propensities into a probability distribution. Category propensities depend on a vector of person n's scores on D measured substantive trait dimensions ($\boldsymbol{\theta}_n$), person n's faking score η_n, a $(K + 1)$-dimensional vector of class-specific item-category intercepts $\boldsymbol{\gamma}_{i(\mathrm{SF})}$, a D-dimensional vector of item slopes of substantive traits $\boldsymbol{\alpha}_{i\mathrm{S}}$, an item slope of faking $\alpha_{i\mathrm{F}}$, and a $((D + 1) \times (K + 1))$-dimensional matrix of scoring weights $\boldsymbol{S}_i$. Substantive trait scores θ_{nd} represent the level of person n on the trait of interest d, faking scores η_n indicate the degree to which person n aligns his or her responses with the items' desirability characteristics.[1] Item slopes α_{id} represent the relation between item i and dimension d, whereas scoring weights s_{idk} reflect the item-specific relation between category k and dimension d. Because of this property of scoring weights, scoring weights are used to specify the to-be-modeled latent dimensions based on theoretical assumptions. For a substantive trait that is measured by a particular item, a scoring weight vector of evenly-spaced integers is specified, as higher response categories should go along with higher substantive trait levels. In contrast, for substantive traits not measured by an item, scoring weights are set to 0. Analogously, scoring weights are specified for a faking dimension. Since scoring weights code relationships between categories and dimensions, scoring weights of faking can be specified such that they reflect the desirability characteristics of individual items (Seitz et al., 2024; Seitz, Spengler, & Meiser, 2025). To set scoring weights of faking in empirical contexts, one possibility is to use desirability ratings of item-category combinations from a pilot study (see Figure 2). Note that,

[1] Note that, as with all latent variables, faking scores are a normative characterization of faking behavior (see Bolt & Meng, 2025). That is, negative faking scores do not indicate "faking bad" in the sense of socially *un*desirable responding, nor does a faking score of 0 indicate the absence of faking. Instead, faking scores should be interpreted as values on a latent continuum, where 0 is usually the latent mean that is set for model identification. Negative faking scores simply reflect a faking level that is below average, whereas the absence of faking is rather implied by a faking slope of 0, which is the case in the model's "S-only class".

because scoring weights are item- and category-specific, this modeling approach of faking allows accounting for various kinds of item desirability characteristics within a test. Thus, not only monotonic desirability trajectories (where desirability increases/decreases monotonically with higher categories; see Figure 2a) but also nonmonotonic ones (where the category of highest desirability is not an extreme category; see Figures 2b and 2c) can be modeled (cf. Kuncel & Tellegen, 2009).

Figure 2

Desirability Trajectories of Three Exemplary Items From the Empirical Demonstration



Note. Mean desirability ratings are based on $N = 74$ participants (data from Seitz, Alagöz, & Meiser, 2025, Pilot Study 2). Error bars represent the standard error of the mean.

In both the "S-only class" and "F-only class", theory-motivated constraints are imposed on item slope parameters (Alagöz & Meiser, 2024; Seitz, Alagöz, & Meiser, 2025). Because responses are not influenced by faking in the "S-only class" ($\zeta_{ni} = 0$), slopes of faking are set to 0 in this class, such that category propensities do not depend on faking scores:

$$p(y_{ni} = k \mid \zeta_{ni} = 0) = \text{softmax}_k\left(\sum_{d=1}^{D} \alpha_{i\text{S}_d} s_{i\text{S}_d k} \theta_{nd} + \gamma_{ik(\text{S})}\right) . \tag{2}$$

This model is equivalent to a multidimensional generalized partial credit model (MGPCM; Muraki, 1992). By contrast, in the "F-only class" ($\zeta_{ni} = 2$), slopes of substantive traits are set to 0. Category propensities are thus independent of substantive trait scores, implying a unidimensional nominal response model (NRM; Bock, 1972) with specified scoring weights of faking:

$$p(y_{ni} = k \mid \zeta_{ni} = 2) = \text{softmax}_k\left(\alpha_{i\text{F}} s_{i\text{F}k} \eta_n + \gamma_{ik(\text{F})}\right) . \tag{3}$$

Note that non-fixed slopes in this model are class-invariant, whereas item-category intercepts are class-specific. This has the purpose of measuring the same latent variables across classes, while at the same time being able to capture different item response distributions in the three classes. For model identification, the intercept of the first category is fixed to 0 for all items in all classes.

Item Response Time Models

Research has repeatedly demonstrated RT differences associated with response sets like honest responding, responding under conditions of heightened desirability concerns, or instructed faking (e.g., Holden et al., 1992; Holtgraves, 2004). To utilize this information in our mixture modeling approach, we specify different RT models associated with the different response strategies (see Ulitzsch et al., 2020; Ulitzsch, Pohl, et al., 2022, for similar approaches). Specifically, RTs t_{ni} are modeled through log-normal models featuring a person speed parameter φ_n and item time intensity parameters δ_i (van der Linden, 2006). Speed is part of a multivariate normal distribution together with the other person parameters of the model (van der Linden, 2007). Time intensities reflect the predicted log-RTs for test-takers with a speed score of 0 (which represents the latent mean in our case).

To capture mean differences in RTs between classes, item time intensities are class-specific. Regarding the question of whether honest responding or faking takes longer, the literature is inconsistent. Some studies found faking to be associated with reduced RTs (e.g., Holden et al., 1992; Holden, 1995; Hsu et al., 1989), other studies found faking to increase RTs (e.g., Fine & Pirak, 2016; Holtgraves, 2004; Walczyk et al., 2003). Hence, no constraints are put on time intensities in the "S-only class" ($\zeta_{ni} = 0$) and "F-only class" ($\zeta_{ni} = 2$):

$$\ln(t_{ni} \mid \zeta_{ni} = 0) \sim N(\delta_{i(\mathrm{S})} - \nu_\varphi \varphi_n, \varsigma^2_{(\mathrm{S})}) \,, \tag{4}$$

$$\ln(t_{ni} \mid \zeta_{ni} = 2) \sim N(\delta_{i(\mathrm{F})} - \nu_\varphi \varphi_n, \varsigma^2_{(\mathrm{F})}) \,. \tag{5}$$

The parameter ν_φ is an item-invariant speed slope, $\varsigma^2_{(\mathrm{S})}$ and $\varsigma^2_{(\mathrm{F})}$ denote class-specific residual variances of log-RTs. In the "S&F class", time intensities are constrained to be a function of "S-only class" time intensities $\delta_{i(\mathrm{S})}$ and “F-only class” time intensities $\delta_{i(\mathrm{F})}$, as responses in this class are based on substantive traits *and* entail response editing according to the items’ desirability characteristics:

$$\ln(t_{ni} \mid \zeta_{ni} = 1) \sim N(\delta_{i(\mathrm{S})} + \lambda\delta_{i(\mathrm{F})} - \nu_{\varphi}\varphi_n, \varsigma^2_{(\mathrm{SF})}) , \tag{6}$$

with $\varsigma^2_{(\mathrm{SF})}$ denoting the class-specific residual variance of log-RTs, and λ being a proportionality constant on $\delta_{i(\mathrm{F})}$. Because both response processes are involved in the “S&F class”, we expect mean RTs in this class to be longer than mean RTs in both the “S-only class” and “F-only class” (Walczyk et al., 2003). To encode this assumption, λ is constrained to be positive. It hence reflects the extent to which faking along with substantive trait responding increases RTs compared to pure substantive trait responding.[2]

Note that, in the proposed model, RTs are modeled as indicators of class membership as opposed to predictors of class membership (see Nagy & Ulitzsch, 2022; cf. Ulitzsch et al., 2020; Ulitzsch, Pohl, et al., 2022). That is, RTs are reflections of strategy use and contribute to the classification of responses by their relative typicality for the respective class. Since RT distributions are allowed to overlap, classification is not based on fixed RT cutoffs. While “S&F class” responses are assumed to take on average longer than their single-process counterparts, the overlap of RT distributions allows also short responses to be classified as “S&F class” responses, as well as also long responses to be classified as “S-only class” or “F-only class” responses.

Latent Response Model

Test-takers’ latent class memberships ζ_{ni} are not observable. However, class memberships determine the measurement model of individual responses and RTs, and thus represent latent response variables (see Maris, 1995; cf. Ulitzsch et al., 2020; Ulitzsch, Pohl, et al., 2022; Ulitzsch, Yildirim-Erbasli, et al., 2022). To model item-wise class membership, we treat the three response strategies as

[2] To corroborate the assumption of longest RTs in the “S&F class”, we conducted a pre-study in which we assessed RT behavior associated with the three modeled response strategies (see Appendix). In line with theoretical considerations, RTs in the pre-study were indeed longest when participants responded according to substantive traits and desirability. Also, RTs were shorter when participants responded only according to desirability than when they responded only according to substantive traits.

ordinal, with the “F-only class” reflecting the most extreme self-presentation strategy, the “S&F class” reflecting an intermediate self-presentation strategy, and the “S-only class” reflecting the least pronounced self-presentation strategy. Another person parameter, strategy inclination ψ_n, as well as item-class intercepts β_{ic} determine the propensity of being a member of class c on item i. A softmax function transforms the propensities into probabilities, which are the mixing proportions in the proposed person-by-item mixture model:

$$p(\zeta_{ni} = c) = \mathrm{softmax}_c\left(\nu_{\psi} c \psi_n + \beta_{ic}\right)\,. \tag{7}$$

This model represents a partial credit model (PCM; Masters, 1982) with an item-invariant strategy inclination slope ν_{ψ}. Conceptually, strategy inclination can be interpreted as the tendency of a person to use a more pronounced self-presentation strategy, that is, prefer the “F-only class” over the “S&F class” and the “S&F class” over the “S-only class”. Item-class intercepts can be interpreted as item easiness parameters associated with a particular strategy. For identification, the intercept of the first class (“S-only class”) is fixed to 0 for all items.

The $D + 3$ person parameters of the proposed model are assumed to follow a joint multivariate normal distribution with expectation vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{\theta_1} \\ \vdots \\ \mu_{\theta_d} \\ \vdots \\ \mu_{\theta_D} \\ \mu_{\eta} \\ \mu_{\varphi} \\ \mu_{\psi} \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2_{\theta_1} & & & & & & & \\ \vdots & \ddots & & & & & & \\ \sigma_{\theta_1\theta_d} & \vdots & \sigma^2_{\theta_d} & & & & & \\ \vdots & \vdots & \vdots & \ddots & & & & \\ \sigma_{\theta_1\theta_D} & \vdots & \sigma_{\theta_d\theta_D} & \vdots & \sigma^2_{\theta_D} & & & \\ \sigma_{\theta_1\eta} & \cdots & \sigma_{\theta_d\eta} & \cdots & \sigma_{\theta_D\eta} & \sigma^2_{\eta} & & \\ \sigma_{\theta_1\varphi} & \cdots & \sigma_{\theta_d\varphi} & \cdots & \sigma_{\theta_D\varphi} & \sigma_{\eta\varphi} & \sigma^2_{\varphi} & \\ \sigma_{\theta_1\psi} & \cdots & \sigma_{\theta_d\psi} & \cdots & \sigma_{\theta_D\psi} & \sigma_{\eta\psi} & \sigma_{\varphi\psi} & \sigma^2_{\psi} \end{pmatrix}. \tag{8}$$

To identify the scale of latent variables, latent means are set to 0 and latent variances to 1, such that latent covariances represent latent correlations. Assuming conditional independence of item responses and RTs, the model’s joint likelihood of the data, marginalized over the three latent classes, can be denoted as

$$\mathcal{L} = \prod_{n=1}^{N}\prod_{i=1}^{I} \begin{array}{c} (p(\zeta_{ni} = 0 \mid \psi_n, \nu_\psi, \beta_{ic})\, p(y_{ni} \mid \boldsymbol{\theta}_n, \boldsymbol{\alpha}_{i\mathrm{S}}, \boldsymbol{S}_i, \boldsymbol{\gamma}_{i(\mathrm{S})})\, f(t_{ni} \mid \varphi_n, \nu_\varphi, \delta_{i(\mathrm{S})}, \varsigma^2_{(\mathrm{S})}) + \\ p(\zeta_{ni} = 1 \mid \psi_n, \nu_\psi, \beta_{ic})\, p(y_{ni} \mid \boldsymbol{\theta}_n, \eta_n, \boldsymbol{\alpha}_{i\mathrm{S}}, \alpha_{i\mathrm{F}}, \boldsymbol{S}_i, \boldsymbol{\gamma}_{i(\mathrm{SF})})\, f(t_{ni} \mid \varphi_n, \nu_\varphi, \delta_{i(\mathrm{S})}, \delta_{i(\mathrm{F})}, \lambda, \varsigma^2_{(\mathrm{SF})}) + \\ p(\zeta_{ni} = 2 \mid \psi_n, \nu_\psi, \beta_{ic})\, p(y_{ni} \mid \eta_n, \alpha_{i\mathrm{F}}, \boldsymbol{S}_i, \boldsymbol{\gamma}_{i(\mathrm{F})})\, f(t_{ni} \mid \varphi_n, \nu_\varphi, \delta_{i(\mathrm{F})}, \varsigma^2_{(\mathrm{F})})) \end{array}, \quad (9)$$

where N is the total number of persons, I is the total number of items, and $f(\ \dots)$ is the log-normal density of RTs.

Model Estimation

The proposed person-by-item mixture model can be estimated with a Bayesian Markov chain Monte Carlo (MCMC) procedure. We used the software *JAGS* (version 4.3.2; Plummer, 2017) for model estimation, accessed through the *R* environment (version 4.4.3) using the package *runjags* (Denwood, 2016). To process MCMC outputs, we employed the packages *coda* (Plummer et al., 2006) and *MCMCvis* (Youngflesh, 2018). The JAGS syntax as well as R code for estimating the model can be found at https://osf.io/crmv4/?view_only=db3eb1f7139742f48ee01887363b5e4a.

We used the following prior distributions for the different model parameters: Slope parameters were drawn from positively-truncated normal priors ($\alpha_{i\mathrm{S}_d}, \alpha_{i\mathrm{F}}, \nu_\varphi, \nu_\psi \sim N^+(0, 2^2)$). Class-specific item-category intercepts in the item response model components were sampled from uncensored normal priors ($\gamma_{ik(\mathrm{S})}, \gamma_{ik(\mathrm{SF})}, \gamma_{ik(\mathrm{F})} \sim N(0, 4^2)$), with the intercept of the first category fixed to 0 for model identification. The same prior was used for item-class intercepts in the latent response model component ($\beta_{ic} \sim N(0, 4^2)$, intercept of the first class fixed to 0 for identification). For "S-only class" and "F-only class" item time intensities in the item RT model components, normal priors were used as well, with the empirical mean of log-RTs across persons and items serving as prior center ($\delta_{i(\mathrm{S})}, \delta_{i(\mathrm{F})} \sim N(\overline{\ln(t_{ni})}, 1^2)$). Since the "S&F class" proportionality constant was constrained to be greater than 0 but was not expected to take on large values, a positively-truncated normal prior with a rather small variance was employed for this parameter ($\lambda \sim N^+(0, 0.5^2)$). For residual standard deviations of log-RTs, standard half-Cauchy priors were used, which are positively-truncated central *t*-distributions with 1 degree of freedom ($\varsigma_{(\mathrm{S})}, \varsigma_{(\mathrm{SF})}, \varsigma_{(\mathrm{F})} \sim t^+(1)$). The prior for person parameters was a multivariate normal distribution ($(\boldsymbol{\theta}_n, \eta_n, \varphi_n, \psi_n) \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma}$ as a variance-covariance matrix with unit variances for identifying the scale of latent variables. Covariances, which reflected correlations, had uniform priors of $U(-1, 1)$. To sample class membership ζ_{ni}, values were drawn from a categorical distribution $Cat(\boldsymbol{\pi}_{ni})$, with $\boldsymbol{\pi}_{ni}$ being a vector of person- and item-specific class probabilities resulting from Equation 7 at the respective MCMC iteration.

We considered means of posterior distributions as point estimates of model parameters. Additionally, we derived posterior class probabilities and class proportions estimates. Since class probabilities are a direct function of the parameters from the latent response model, posterior class probabilities can be computed by plugging the samples of the latent response model parameters into Equation 7 for each non-discarded MCMC iteration before aggregating across iterations. To get class proportion estimates, mean class probabilities across persons and iterations can be calculated, either per item or aggregated across items. It is worth noting that class labels in the proposed model are not arbitrary since the measurement models of the three classes are explicitly specified. This confirmatory specification of latent classes avoids the issue of label switching, which is often encountered when estimating mixture models in a Bayesian framework (Jasra et al., 2005).

Differences to Related Faking Mixture Models

As elaborated above, most faking models do not account for a faking process that varies between items. However, there are some models that do allow for inter-item variation in how item responses are (not) affected by substantive traits and faking. Böckenholt (2014), for instance, presented a model of motivated misreports that is characterized by possible response editing before providing answers to sensitive survey questions (see also Leng et al., 2020). This model includes for every item a dichotomous latent class variable indicating if a test-taker edits his or her retrieved response. In case of editing, a transition function models the selection of response categories that are more desirable than the category that would have been chosen without editing. Though theoretically appealing, the model is limited to only one substantive trait dimension. Also, response editing is restricted to go in the direction of high (low) categories if the measured substantive trait is generally desirable (undesirable).

Another model allowing for switches between response strategies over the course of the test is Brown and Böckenholt's (2022) model of intermittent faking. This model assumes that item responses are mixtures of honest and faked responses, with honest ("real") responses influenced by substantive traits and faked ("ideal") responses influenced by a faking factor. Whether responses are honest or faked is modeled through an editing factor and item characteristics. This model is not limited to a single substantive trait, however, its frequentist estimation curbs the number of dimensions that can practically be modeled. Also, the model adopts a structural equation modeling (SEM) approach, which was empirically demonstrated using total scale scores as indicator variables and not individual items. In addition, the possibility of nonlinear and nonmonotonic item desirability trajectories (Kuncel & Tellegen, 2009; Seitz et al., 2024; see Figure 2) is not directly taken into account. Furthermore, the model does not include a class where substantive traits *and* faking influence item responses. However, such a class is conceivable because studies have found faking to operate at the editing stage of item responding and not at the retrieval stage (Holtgraves, 2004; Sudman et al., 1996; Walczyk et al., 2003). Robie et al. (2007) also showed that a considerable share of test-takers in high-stakes assessments refer to both their true personality and the ideal applicant when giving item responses.

Along with addressing several limitations of related mixture models of faking (modeling multiple substantive traits, nonmonotonic faking effects, and an "S&F class"), the proposed model also incorporates item-level RTs. As elaborated above, this can provide valuable insights into the cognitive processes associated with faking. The use of collateral data can also be expected to facilitate the estimation of such a person-by-item mixture model, where the information for class assignment is very sparse compared to a mixture model with constant class membership per person.

Simulation Study

To evaluate the proposed model under realistic conditions, we conducted a simulation study. The purpose of this study was twofold: a) to examine the model's ability to accurately recover parameters and b) to compare it to the performance of less complex models.

Data Generation and Fitted Models

For emulating realistic conditions, data-generating values for item responses and item RTs were in line with parameter estimates from the empirical demonstration below. Resembling our empirical dataset, data from a test measuring 3 substantive traits with 10 items each on a 7-point Likert scale were simulated using the R packages *MASS* (Venables & Ripley, 2002) and *extraDistr* (Wolodzko, 2023). For the simulation study, a sample size of $N = 500$ was considered. 50 independent replications were performed. Further details on the data generation procedure can be found in the Supplement, which can be accessed here.

Each simulated dataset was analyzed using the proposed person-by-item mixture model including RTs. In addition, four less complex models were fitted to every dataset: a person-by-item mixture model not accounting for RTs, a person mixture model accounting for RTs, a person mixture model not accounting for RTs, as well as a non-mixture model. Comparing models with and without RTs allowed investigating how the inclusion of RTs improves model estimation in terms of parameter recovery and class separation. Comparing person-by-item and person mixture models allowed investigating the consequences of disregarding varying strategy use across items. Comparing mixture and non-mixture models allowed examining the effect of assuming a single response strategy across persons and items.

The non-mixture model was an MNRM accounting for substantive traits and faking with specified scoring weights (see Equation 1). All mixture models included the three described latent classes, however, the two person mixture models did not have the latent response model component. Instead, latent class membership in the person mixture models was only a person variable sampled from a categorical distribution $Cat(\boldsymbol{\pi})$ with a flat Dirichlet hyperprior ($\boldsymbol{\pi} \sim Dir(\mathbf{1})$). The mixture models without RTs were equivalent to their RT counterparts, with the exception of not having the item RT model components. All model syntaxes are available at https://osf.io/crmv4/?view_only=db3eb1f7139742f48ee01887363b5e4a. Scoring weights were specified as in the data generation. For the Bayesian estimation of each model, 4 parallel MCMC chains were run with 4000 burnin and 10000 regular iterations.[3] Model convergence was assessed based on $\hat{R}$ (Gelman & Rubin, 1992), with a model considered as converged if all model parameters had $\hat{R}$ values below 1.1.

Results of the Simulation Study

The person-by-item mixture model with RTs converged in 49 out of 50 replications (98%). The person-by-item mixture model without RTs, however, only converged in 37 replications (74%). The person mixture model with RTs converged in 39 replications (78%), the person mixture model without RTs in 48 replications (96%). The non-mixture model converged in all 50 replications (100%).

We analyzed parameter recovery based on converged models. In particular, we looked at bias to examine systematic over- or underestimation of parameters as well as at root mean square error (RMSE) to investigate estimation accuracy. For the recovery of person parameters, we considered the correlation between estimated and true parameters. Results are displayed in Figures 3 to 8. Overall, the data-generating person-by-item mixture model with RTs estimated parameters with negligible bias. Only for item-category intercepts, the model yielded slightly negatively biased estimates (corresponding to an average underestimation of 5-10% compared to the data-generating values). In contrast, all of the other models produced biased estimates for most of the model parameters. Regarding RMSE, too, the less complex models yielded for all parameters worse results than the person-by-item mixture model with RTs. Even though such a result pattern can in principle be expected since the data were generated based

[3] Mixture models can be challenging to estimate because of multimodal likelihood functions (McLachlan & Peel, 2000), which make single MCMC chains susceptible to converging to local solutions. In order to alleviate this problem, we employed an estimation strategy that optimizes initial values before the actual estimation (cf. O'Hagan et al., 2012). In the simulation study, this was done by running 12 (30 in the empirical demonstration) parallel chains (each with 2000 burnin and 2000 regular iterations (3000+3000 in the empirical demonstration)) with unsystematic initial values and taking the parameter estimates from the chain yielding the highest model likelihood as initial values for the actual model estimation. The actual estimation then featured the reported numbers of chains and iterations, with random noise added to the initial values of each chain to avoid that every chain started with the exact same set of values. Details can be retrieved at https://osf.io/crmv4/?view_only=db3eb1f7139742f48ee01887363b5e4a.

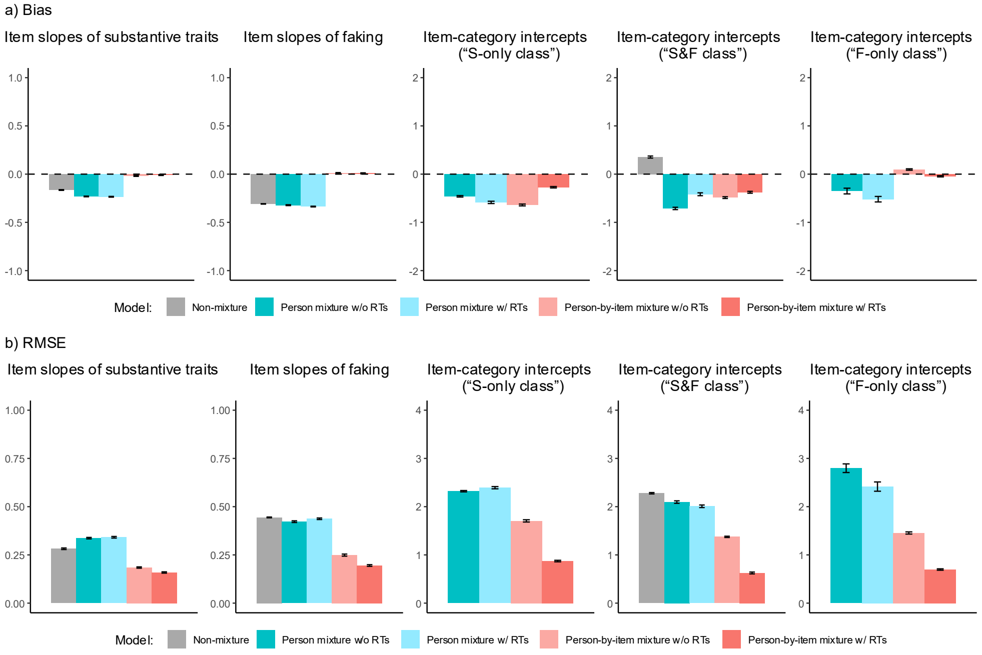
on a person-by-item mixture population, the results do indicate that conclusions drawn from models that do not capture the full data-generating process are indeed considerably less accurate. For instance, RMSE for RT model parameters (Figure 4b) and latent response model parameters (Figure 5b) was much larger in the corresponding person mixture model and person-by-item mixture model without RTs than in the full model. In addition, whereas both person-by-item mixture models yielded unbiased and fairly accurate estimates of class proportions, both person mixture models severely overestimated the overall proportion of the "S&F class" and underestimated the "F-only class" proportion (Figure 6). The person mixture model without RTs additionally overestimated the proportion of the "S-only class". Regarding person parameters, the less complex models also yielded poorer estimates than the data-generating person-by-item mixture model with RTs (Figure 8). Effects were most pronounced for faking scores, where correlations between estimated and true person parameters differed considerably between models. For substantive trait, speed, and strategy inclination scores, the person-by-item mixture model with RTs yielded the highest correlations as well, but the less complex models did not perform much worse.

Furthermore, we examined the mixture models' classification accuracy based on a modal class assignment rule (i.e., assignment to the class with the highest posterior class probability; Dias & Vermunt, 2008). We hereby looked at item-level hit rates, which indicate percentages of correct class assignments of individual item responses. The person-by-item mixture model with RTs yielded a mean hit rate of 63.3%, whereas the person-by-item mixture model without RTs only afforded a mean hit rate of 58.5%. Classification accuracy in the person mixture model with RTs (hit rate: 51.0%) and without RTs (hit rate: 51.7%) was even lower, which is conceivable given that a person's class membership was not constant in the data generation. The size of the observed hit rates will be put into context in the Discussion section.[4]

[4] We also conducted two additional simulations with special-case population models: a) a person mixture population model in which class membership of a person was constant across items and b) a non-mixture population model in which the whole sample belonged to a single class. Results showed that all mixture models can afford much higher classification accuracy in such conditions (hit rates: 92.7%-100%). For person-by-item mixture models, this can be achieved by approximating constant class membership per person or the entire sample through the parameters of the latent response model component. Nevertheless, the person-by-item mixture models yielded slightly less accurate parameter estimates compared to the respective population model, suggesting that the proposed model can be prone to overfitting. Details on the additional simulations are reported in the Supplement (to be accessed via this link).

Figure 3

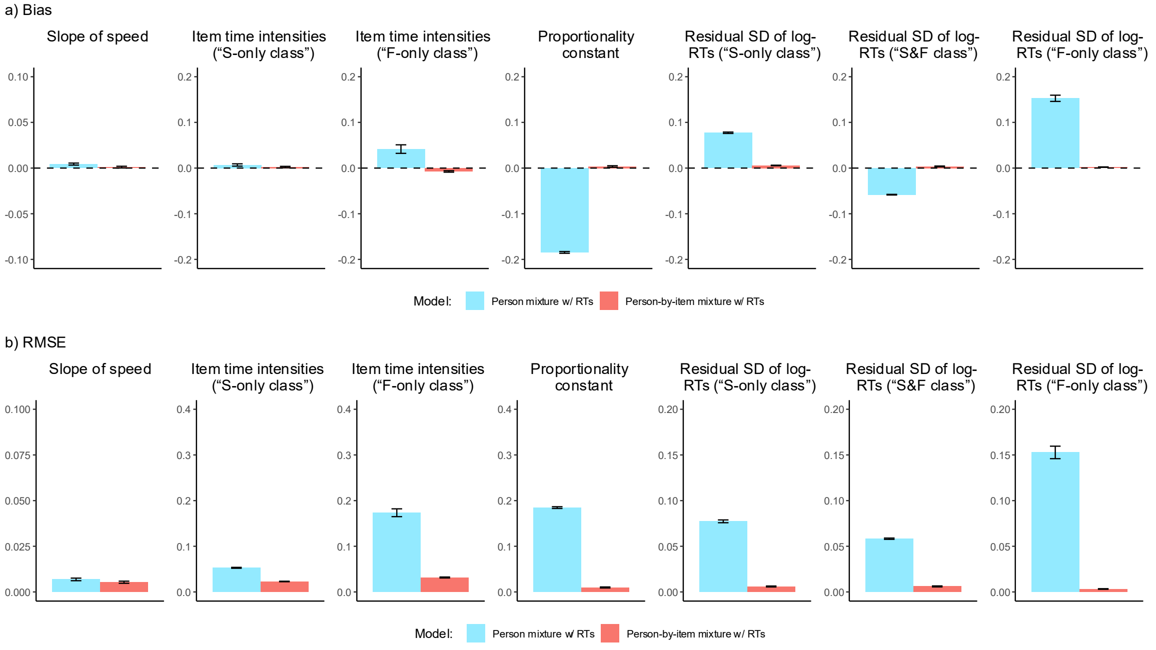
Simulation Study: Recovery of Item Response Model Parameters



Note. Values reflect the mean bias (Panel a) or root mean square error (RMSE; Panel b) of estimated parameters across replications. For mixture models, the respective class-specific intercept estimates are considered. Error bars represent the standard error of the mean. RT = response time.

Figure 4

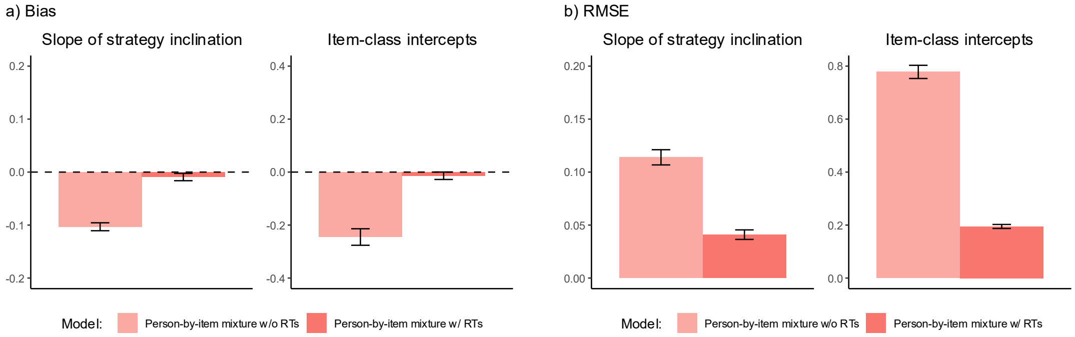
Simulation Study: Recovery of Item Response Time Model Parameters



Note. Values reflect the mean bias (Panel a) or root mean square error (RMSE; Panel b) of estimated parameters across replications. Error bars represent the standard error of the mean. SD = standard deviation; RT = response time.

Figure 5

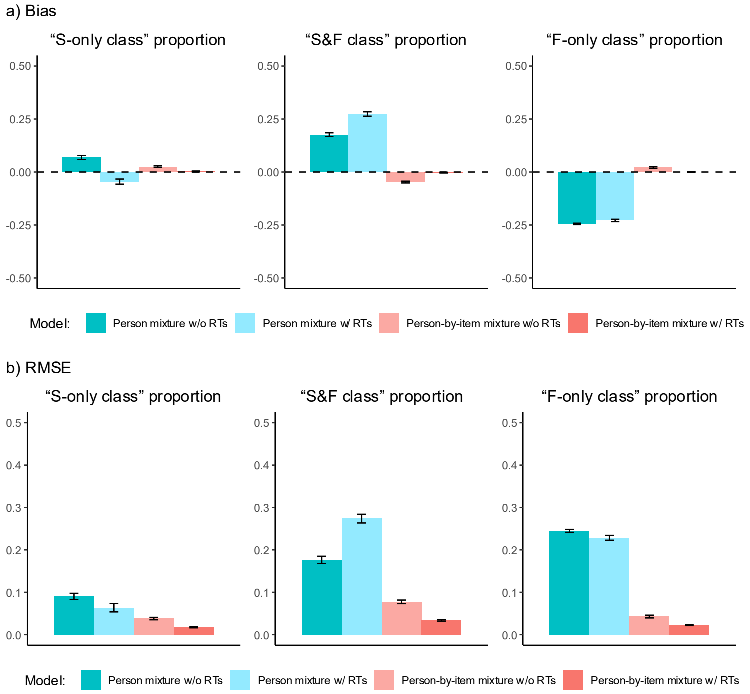
Simulation Study: Recovery of Latent Response Model Parameters



Note. Values reflect the mean bias (Panel a) or root mean square error (RMSE; Panel b) of estimated parameters across replications. Error bars represent the standard error of the mean. RT = response time.

Figure 6

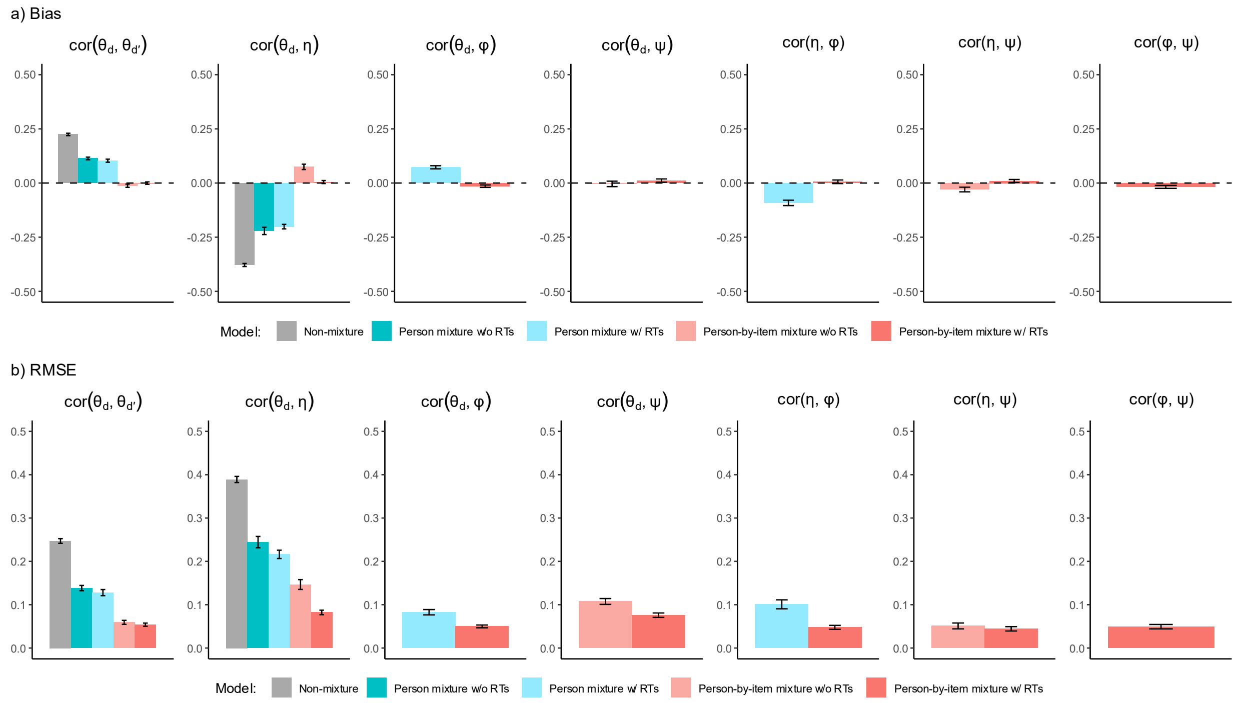
Simulation Study: Recovery of Class Proportions



Note. Values reflect the mean bias (Panel a) or root mean square error (RMSE; Panel b) of estimated parameters across replications. Class proportions are aggregated proportions across items. Error bars represent the standard error of the mean. RT = response time.

Figure 7

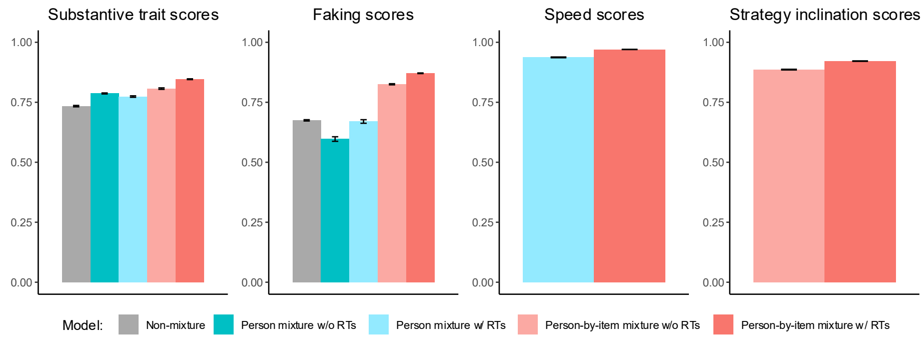
Simulation Study: Recovery of Latent Correlations



Note. Values reflect the mean bias (Panel a) or root mean square error (RMSE; Panel b) of estimated parameters across replications. Results for substantive traits are aggregated across the three substantive traits. Error bars represent the standard error of the mean. θ_d = substantive trait d; η = faking; φ = speed; ψ = strategy inclination; RT = response time.

Figure 8

Simulation Study: Recovery of Person Parameters



Note. Values reflect the mean correlations (using Fisher's *z*-transformation) between estimated and true parameters across replications. Results for substantive traits are aggregated across the three substantive traits. Error bars represent the standard error of the mean.

Empirical Demonstration

To empirically demonstrate the proposed model, we analyzed data from an actual high-stakes job application context. Along with technical aspects like convergence and model fit in empirical settings, the empirical demonstration allowed investigating the consistency of test-takers' response strategy use across items, class proportions on the level of items and the entire test, as well as RT differences associated with the three response strategies.

Dataset

The data for the empirical demonstration were made available by a Germany-based testing company that develops psychological measurement tools for personnel selection. The dataset included $N = 1824$ test-takers who had taken a personality test as part of their application for a police officer traineeship at a German police department between 2023 and 2024.[5] The sample comprised 70.0% male and 30.0% female test-takers with a mean age of $M = 21.02$ years ($SD = 4.61$, $range = [15, 39]$). Along with item responses, the dataset contained item-level RTs (measured down to milliseconds). Exact timing of item-level response latencies was possible because every item had been presented on a separate questionnaire page.

For our empirical demonstration, we modeled item responses and RTs from three substantive trait scales available in the dataset. These were a scale of Emotional Stability (measured with 12 items; Cronbach's $\alpha = .74$, McDonald's $\omega = .73$), a scale of Extraversion (9 items; $\alpha = .67$, $\omega = .65$), and a scale of Conscientiousness (10 items; $\alpha = .77$, $\omega = .79$). Across the substantive trait scales, items appeared in a random order. Responses were given on a 7-point Likert scale (0 = *does not apply at all* to 6 = *applies fully*).

[5] This dataset was also analyzed by Seitz, Alagöz, and Meiser (2025), but their analysis did not involve RTs and was limited to a person mixture model.

Pilot Study to Assess Desirability Values

We set scoring weights of substantive trait dimensions to $(0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)$ and scoring weights of the faking dimension to values representing the social desirability of the items' response categories with regard to the job application context at hand. These desirability values were collected in a pilot study, in which participants rated the desirability of every category of every item regarding the application for a police officer traineeship (details are provided in Seitz, Alagöz, & Meiser, 2025, Pilot Study 2, and can be retrieved at https://osf.io/crmv4/?view_only=db3eb1f7139742f48ee01887363b5e4a). Participants should thereby take the perspective of a person currently applying for such a job and rate desirability accordingly. The resulting mean desirability ratings then served as scoring weights of faking (see Figure 2 for three exemplary items). To achieve a common metric of scoring weights across latent dimensions, we linearly transformed the ratings to a possible range from 0 to 6.

Results of the Empirical Demonstration

Like in the simulation study, we fitted the person-by-item mixture model both with and without RTs as well as the person mixture model with and without RTs. We also fitted the three non-mixture models representing the measurement models of three modeled response strategies (Equations 1 to 3), that is, an MNRM, an MGPCM, and a unidimensional NRM with specified scoring weights of faking. We again used JAGS via the R environment for model estimation. In the empirical analysis, we estimated each model by running 12 parallel MCMC chains with 15000 iterations following a 5000-iteration burnin phase.

Model Convergence and Model Fit

We checked model convergence based on $\hat{R}$. For the person-by-item mixture model with RTs, all model parameters had $\hat{R}$ values below 1.1. A visual inspection of trace plots also showed well-mixed MCMC chains of the different model parameters. However, both the person-by-item mixture model without RTs and the person mixture model with RTs did not converge with all $\hat{R}$ values below 1.1, which is in line with the reduced convergence rates of these two models in the simulation study. For the person-by-item mixture model without RTs, the non-convergence was mainly due to several item-class intercepts showing poorly-mixed trace plots. For the person mixture model with RTs, the lack of convergence mainly stemmed from single item time intensities failing to converge. In contrast, the person mixture model without RTs as well as the three non-mixture models did converge with all $\hat{R}$ values below 1.1. In the following, we report results for the converged models only.

Table 1 shows model fit indices. The person-by-item mixture model with RTs yielded by far the highest log-likelihood.[6] Crucially, this model was also selected by the widely applicable information criterion (WAIC; Watanabe, 2010) and the leave-one-out information criterion (LOOIC; Vehtari et al., 2017), indicating that the person-by-item mixture model with RTs yielded a better compromise between fit and parsimony than the other models. To quantify absolute fit, we applied posterior predictive model checking (PPMC; Sinharay et al., 2006), which is a technique that entails simulating data from the posterior distribution of model parameters and evaluating how the simulated data aligns with the observed data. Regarding item responses, we considered the standardized root mean square residual (SRMR), which indicates the misfit between model-implied and observed item intercorrelations. Compared to the other models, the SRMR of the person-by-item mixture model with RTs was smallest (.054), indicating that this model fit the item responses best. Regarding item RTs, we considered posterior predictive *p*-values (PPP) with respect to the discrepancy between model-implied and observed item means of log-RTs. Across items, PPPs ranged from .182 to .554 with a mean of .382, indicating that the RTs predicted by the model were not systematically higher or lower than the empirical RTs.[7]

[6] Note that likelihood-based comparisons (i.e., log-likelihood differences, WAIC, and LOOIC) between models accounting versus not accounting for RTs are only warranted if the likelihood is computed with respect to item responses only. If the likelihood is computed with respect to the full modeled data, likelihoods of the different models are not comparable because models not accounting for RTs only comprise item response data and models accounting for RTs additionally comprise item RT data.

[7] In particular, item responses and item RTs were simulated for each MCMC iteration after the burnin phase using the respective model equation and sampled parameter values in the given iteration. Regarding item response data, the SRMR of a model was calculated by computing the root mean squared deviation between item intercorrelations in the simulated versus observed data per iteration before averaging across iterations. Regarding item RT data, item-wise PPPs were computed as percentages of simulated log-RT item means being greater than or equal to the corresponding observed item mean across iterations.

Table 2 contains the latent correlations between dimensions estimated in the person-by-item mixture model with RTs. Most correlations were estimated credibly different from 0, but were not large in size. Note, for instance, the estimated correlation between faking and strategy inclination. This correlation of .32 indicates a positive association between aligning responses with desirability and using a pronounced self-presentation strategy, however, it also speaks against a redundancy of the two dimensions.

Table 1

Empirical Demonstration: Model Fit Indices of Converged Models

		Information criterion		PPMC	
Model	LL	WAIC	LOOIC	SRMR	RT-PPP
Non-mixture model (NRM)	–77966.4	157317.9	158261.5	.142	
Non-mixture model (MGPCM)	–71507.7	146724.0	149050.3	.101	
Non-mixture model (MNRM)	–68981.6	142372.6	144994.4	.064	
Person mixture model w/o RTs	–67561.1	139844.2	142550.6	.062	
Person-by-item mixture model w/ RTs	**–56557.8** **(–61785.7)**	**130162.4** **(154628.0)**	**132862.1** **(157201.5)**	**.054**	**.382** **[.182, .554]**

Note. $N = 1824$. LL, WAIC, and LOOIC values outside round brackets were calculated with respect to item responses only, values in brackets with respect to the joint data of item responses and item response times (RT). LL = log-likelihood; WAIC = widely applicable information criterion; LOOIC = leave-one-out information criterion; PPMC = posterior predictive model checking; SRMR = standardized root mean square residual (based on PPMC); RT-PPP = mean posterior predictive *p*-value (range in square brackets) with respect to the log-RT item means; NRM = unidimensional nominal response model with specified scoring weights of faking; MGPCM = multidimensional generalized partial credit model; MNRM = multidimensional nominal response model. The best-fitting model is printed in bold.

Table 2

Empirical Demonstration: Estimated Latent Correlations

	θ_1	θ_2	θ_3	η	φ	ψ
Emotional Stability (θ_1)	1					
Extraversion (θ_2)	.20 [.13, .27]	1				
Conscientiousness (θ_3)	.24 [.17, .31]	.36 [.29, .43]	1			
Faking (η)	−.23 [−.33, −.13]	−.01 [−.12, .11]	−.24 [−.35, −.14]	1		
Speed (φ)	.11 [.06, .17]	.09 [.03, .15]	.08 [.02, .13]	.01 [−.07, .09]	1	
Strategy inclination (ψ)	−.06 [−.13, .02]	.14 [.06, .21]	−.31 [−.38, −.24]	.32 [.22, .41]	−.00 [−.05, .05]	1

Note. $N = 1824$. Values reflect the estimated latent correlations (95% credible intervals in brackets) in the person-by-item mixture model with response times.

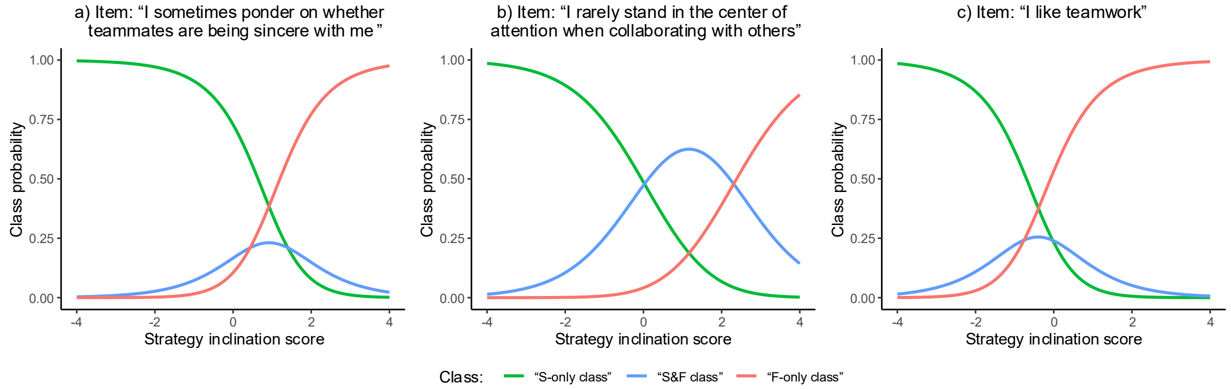
Class Proportions and Class-Specific Item Response Distributions

With an estimated strategy inclination slope of $\widehat{\nu_\psi} = 1.05$ (95% credible interval (CrI): [1.00, 1.11]) and considerable variation in estimated item-class intercepts, the latent response model of the person-by-item mixture model with RTs implied class memberships that differed between both persons and items. The estimated item-class intercepts also implied class proportions that varied substantially from item to item. “S-only class” proportions ranged from 28.9% to 65.6%, “S&F class” proportions from 9.3% to 44.2%, and “F-only class” proportions from 8.0% to 51.6%. Figure 9 displays class probabilities as a function of strategy inclination scores for the three test items with the largest proportion of the “S-only class”, “S&F class”, and “F-only class”, respectively. Across all items, the person-by-item mixture model with RTs estimated that 48.6% (95% CrI: [47.0%, 50.1%]) of individual item responses stemmed from the “S-only class”, whereas 25.9% (95% CrI: [24.5%, 27.4%]) stemmed from the “S&F class” and 25.5% (95% CrI: [24.7%, 26.2%]) stemmed from the “F-only class”. The person mixture model without RTs estimated a similar class proportion for the “S-only class” (48.7%, 95% CrI: [44.7%, 52.7%]), however, it yielded a considerably larger “S&F class” proportion (47.0%, 95% CrI: [43.2%, 51.0%]) and a considerably smaller “F-only class” proportion (4.2%, 95% CrI: [3.2%, 5.3%]). Note that this pattern is in line with the simulation results above, where overall proportions of the “S&F class” and “F-only class” were strongly biased in person mixture models.

Apart from different class sizes across items, the classes also differed in their response distributions depending on the items’ desirability characteristics (see Figure 10). For items at which the highest response category was most desirable (cf. Figure 2a), mean item responses were highest in the “F-only class” and lowest in the “S-only class”. This effect, though slightly less pronounced, also emerged for items having their category of highest desirability above the scale midpoint but not at the extreme (cf. Figure 2b). However, for items with a highest-desirability category at the scale midpoint (cf. Figure 2c), there were no considerable mean differences in item responses between classes.

Figure 9

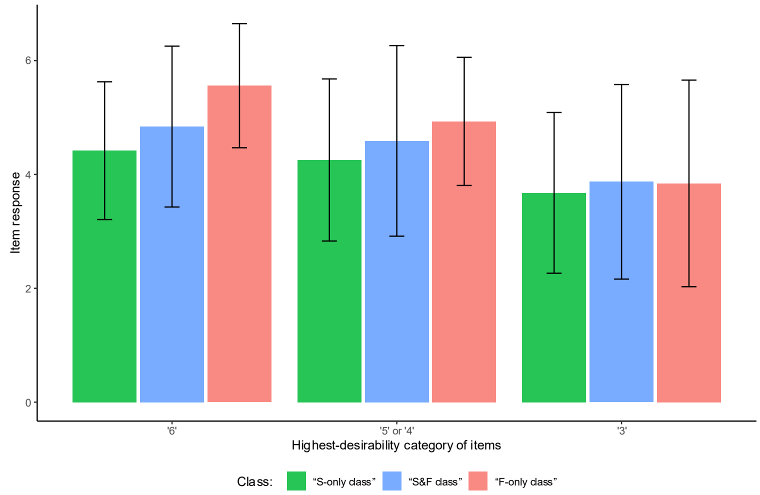
Empirical Demonstration: Latent Response Functions of Three Items



Note. Response functions are based on latent response model estimates from the person-by-item mixture model with response times. Item a had the largest “S-only class” proportion among all test items, item b the largest “S&F class” proportion, and item c the largest “F-only class” proportion.

Figure 10

Empirical Demonstration: Class-Specific Mean Item Responses for Items With Different Desirability Characteristics



Note. Values reflect the mean item response for items with different highest-desirability response categories, split by the class test-takers were assigned to by the person-by-item mixture model with response times. Error bars represent the standard deviation.

Response Time Results

The person-by-item mixture model with RTs also yielded pronounced class differences concerning RTs. Across items, the mean "S-only class" time intensity was $\overline{\widehat{\delta_{.(\mathrm{S})}}} = 1.72$, whereas the mean "F-only class" time intensity was $\overline{\widehat{\delta_{.(\mathrm{F})}}} = 1.53$. This difference of 0.19 units on the log-RT scale was credibly different from 0 (95% CrI: [0.18, 0.20]). At the same time, the estimated "S&F class" proportionality constant was $\hat{\lambda} = 0.23$ (95% CrI: [0.22, 0.24]), which was also substantially above its lower bound of 0. With time intensities of the "S&F class" being a linear function of "S-only class" and "F-only class" time intensities as well as the proportionality constant, the parameter estimates implied a mean "S&F class" time intensity of 2.08. On the level of raw item RTs, these estimates corresponded to median RTs of $Mdn = 5.86$ seconds ($MAD = 2.38$) for responses classified into the "S-only class", $Mdn = 6.50$ seconds ($MAD = 2.89$) for "S&F class" responses, and $Mdn = 4.74$ seconds ($MAD = 2.07$) for "F-only class" responses.

General Discussion

In the current work, we presented a mixture IRT model that allows researchers and practitioners to account for, identify, and investigate different faking-related response strategies. The proposed model assumes that item responses of each test-taker are mixtures of three latent classes (pure substantive trait responding, response editing in the direction of desirability, and pure faking), and entails an RT-based latent response model that classifies item responses on the person-by-item level. RTs are modeled as indicators of class membership. That is, instead of being class membership predictors, RTs are reflections of strategy use and facilitate class separation by their relative typicality for the respective class. The model hence falls into the category of independent latent class IRT models (Nagy & Ulitzsch, 2022).

From a psychometric point of view, the presented person-by-item mixture modeling approach can flexibly account for heterogeneity in faking and thus constitutes an important extension over approaches assuming a constant faking strategy for each test-taker throughout the test or even a single measurement model for all test-takers. From a substantive research perspective, the proposed model can be used to study the response process behind faking in a sophisticated way. For example, differences in estimated item time intensity parameters reflect differences in RT distributions associated with the different response strategies, estimated item-class intercepts reveal which items are especially prone to a particular response strategy, and estimated latent correlations between substantive traits and faking and/or strategy inclination indicate how substantive person attributes are related to different faking tendencies. From an applied testing perspective, the model aids in ensuring the validity of inferences about test-takers, as substantive trait scores are estimated by accounting for potential switches between response strategies over the course of the assessment.

Summary and Discussion of Results

Parameter Recovery

We examined the proposed model in a simulation study with datasets representative of high-stakes assessment data. Along with examining parameter recovery of the proposed model, this study also allowed investigating the consequences of not modeling RTs as well as of assuming a constant response strategy per person or across the entire sample. In general, the proposed person-by-time mixture model with RTs exhibited good parameter recovery in terms of negligible bias and decent accuracy considering the not too large sample size in our simulation study. Also, it was found that models that lack components of the data-generating process, which was based on the results of the empirical demonstration, indeed produce systematically biased and much less accurate parameter estimates than the proposed model. Several other findings stand out and are worth discussing.

First, whereas the person-by-item mixture model with RTs converged in almost all replications (98%), the person-by-item mixture model without RTs converged in only 74% of the replications. Non-convergence of the person-by-item mixture model without RTs was mainly due to several item-class intercepts in the latent response model component failing to converge. This is evidence that RTs have a facilitating effect on the estimation of the presented person-by-item mixture model despite not being direct predictors of class membership (see Nagy & Ulitzsch, 2022). Along with improving convergence, the results also suggest that modeling RTs has a positive effect on the accuracy of parameter estimation. In our simulation study, the size of the effect varied between the different model parameters but was most pronounced for the latent response model parameters.

Second, the recovery of most model parameters differed substantially between person-by-item and person mixture models, with person mixture models yielding biased and less accurate estimates. This is not too surprising considering class membership varied both between persons and items in the data generation. However, it is important to note that person mixture models indeed lead to biased conclusions in data situations in which test-takers employ different response strategies over the course of the test (as was the case in our empirical demonstration). Particularly, class proportions estimated in person mixture models seem to be strongly biased, with the proportion of the “S&F class” systematically overestimated and the proportion of the “F-only class” systematically underestimated.[8]

[8] One should keep in mind that the level class proportion estimates refer to differs between person-by-item and person mixture models. Whereas estimates in person-by-item mixture models refer to class proportions on the level of individual item responses, estimates in person mixture models refer to class proportions on the level of test-takers. Class proportion parameters hence do not have the same conceptual meaning in the two types of mixture models.

Third, the simulation study showed for the proposed model a rather low classification accuracy compared to other recent confirmatory mixture modeling approaches (e.g., Alagöz & Meiser, 2024; Seitz, Alagöz, & Meiser, 2025). Note, however, that the metric considered in the current article were hit rates on the item level (i.e., percentages of correct class assignments on the level of individual item responses), whereas articles studying classification accuracy of person mixture models considered hit rates on the level of test-takers. Taking into account that the information for assigning single item responses to latent classes is very sparse and naturally less reliable than the information for classifying whole response vectors, it is unsurprising that the obtained hit rates were not as high as the hit rates found in articles on person mixture models. Nevertheless, considering the size of the observed hit rates in the simulation study, it is important to emphasize that response strategy classifications of individual item responses should be interpreted with caution. Because of the uncertainty associated with single classifications, it is arguably more advisable to look for general trends regarding the model-implied class memberships, be it for a particular test-taker or for a particular item.

Empirical Results

Seitz, Alagöz, and Meiser (2025) provided evidence for the prevalence of the three modeled response strategies on the person level. The empirical demonstration of the current article allowed investigating the consistency of response strategy use over the course of the assessment. As described above, the estimated parameters in the latent response model component implied class memberships that varied between persons and, crucially, within persons also between items. This empirically justifies the use of the more complex person-by-item mixture modeling approach. Besides, note that the flexibility of such a model also allows for a constant class membership if a person's response pattern suggests that, which was indeed the case for 639 out of the 1824 test-takers (35.0%).

Regarding the overall class proportions, the model estimated that about one-half of individual item responses stemmed from the "S-only class", whereas about one-quarter each stemmed from the "S&F class" and the "F-only class". That is, according to the model, around half of the responses were honest in terms of being solely influenced by substantive traits, whereas the other half were at least partially influenced by faking. These proportions are well in line with the proportions Brown and Böckenholt (2022) found in their model including only two classes (49% honest vs. 51% faked). However, our results suggest that some responses influenced by faking can still contain information on substantive traits and hence do not need to be discarded for the estimation of substantive trait scores.

Another interesting empirical finding concerns the model's RT results, indicating that pure faking is faster than pure substantive trait responding and that a combination of the two processes takes longest. These findings could reconcile some of the conflicting results in the literature regarding the effect of faking on RTs. In particular, our findings suggest that the question of whether faking increases or decreases RTs cannot be answered straightforwardly, but that it depends on the response process underlying a response that involves faking. If test-takers retrieve an honest response before they edit this response according to desirability (i.e., "S&F class"; cf. the theory by Walczyk et al., 2003), the results suggest that this goes along with increased RTs (Fine & Pirak, 2016; Holtgraves, 2004; Walczyk et al., 2003). Such a response set corresponds to the response set triggered by the experimental condition in the study by Holtgraves (2004), who induced faking along with honest responding through a subtle desirability manipulation. If, however, test-takers bypass the retrieval of an honest response and instead just provide a desirable answer (i.e., "F-only class"; cf. the theory by Holden, 1995), the results suggest that this reduces RTs. Such a response set is analogous to response sets induced through mere instructed-faking conditions, in which subjects are simply asked to fake without referring to their true personality (Holden et al., 1992; Holden, 1995; Hsu et al., 1989). Taking this into account, the model's RT results imply a limited potential of simple (i.e., non-model-based) faking detection methods that perform faking classifications based on a single RT threshold (Fine & Pirak, 2016).

Limitations and Future Research Directions

Some limitations and directions for future research need to be mentioned. One aspect concerns the model's assumptions: First, in order for RTs to have a facilitating effect on model estimation and class separation, differences in RT distributions should exist between classes. As RT distributions become less distinct, one can expect the facilitating effect of modeling RTs shown in the simulation study to weaken and eventually disappear (Pokropek, 2016; Ulitzsch et al., 2024). Second, along with separable RT distributions, item response distributions should also exhibit good separability between classes. Future research can explore which questionnaire characteristics are required to achieve this (cf. Uglanova et al., 2025). Third, completion speed is assumed to be constant across all items of the test. This stationarity assumption (van der Linden, 2007), however, is likely to be violated when test-takers become more acquainted with the questionnaire format or become exhausted at later parts of the assessment.

Another aspect concerns the current parameterizations of some of the model components. For instance, the latent response model has been parametrized as a PCM, which assumes the classes to be ordinal. For a more general formulation, the latent response model could in principle also be implemented as an NRM, which would relax the ordinality assumption. However, in initial simulation runs in the model development phase, we found this increased estimation complexity to pose a challenge to model convergence. Different parameter restrictions would probably be necessary in this case. Likewise, future studies could try to estimate time intensities of the “S&F class” in a less restrictive way. Based on theoretical considerations, we constrained “S&F class” time intensities to be a function of “S-only class” and “F-only class” time intensities and restricted the proportionality constant λ to be positive. The proportionality constant indeed turned out to be substantially larger than 0 in the empirical demonstration, but an unconstrained estimation of “S&F class” time intensities would nonetheless be interesting. Moreover, a different parameterization of the “F-only class” could be tested, namely one with a model of stochastic independence. In such a model, faking would only be captured by item-category intercepts, with no variance explained by a common factor such that all variation in this class would be unsystematic.

A further alternative way of parameterizing the model would also be to implement item-level predictors of class membership and/or RTs. This could be achieved by restricting item-class intercepts and/or item time intensities through linear combinations of variables that describe the items, such as item keying, item length, item position, or item content features (see Ulitzsch, Yildirim-Erbasli, et al., 2022). Researchers could use such a parametrization to study item characteristics that make certain response strategies especially likely or have a systematic effect on response latencies. If, for instance, particular item characteristics were found to strongly predict “F-only class” membership, this would be a valuable piece of information for test construction, as it could help to develop instruments that are less susceptible to faking in the first place.

Additionally, it would be important to conduct validation analyses in future studies. From the perspective of applied measurement, it would be particularly relevant to further investigate the validity of the model’s adjustments of substantive trait score estimates, because substantive trait scores are the parameters of primary interest in contexts where test-takers are ranked and selected based on their test scores. In this regard, future research should study whether the adjustments of substantive trait score estimates increase correlations with faking-resistant measures of personality. Ultimately, it would also be interesting to see the effects of the adjustments on the prediction of job-relevant performance outcomes.

Finally, practical issues associated with the presented method of modeling faking are to be mentioned. In particular, the presented approach comes with the limitation of requiring a lot of resources in the form of computation power and time. Fitting the model can demand a substantial amount of working memory and take quite some time depending on sample size and test length. A single estimation in our simulation study, for example, took around twelve hours on a high-performance computer. Also, applying the model requires knowledge in software packages for Bayesian estimation, which may limit its use by applied researchers and practitioners.

Conclusion

In conclusion, the presented RT-based person-by-item mixture model constitutes an appealing approach to modeling faking in high-stakes personality testings. As opposed to most other faking models, it accounts for switches between different faking-related response strategies over the course of the assessment. It thereby makes use of additional behavioral data by modeling strategy-specific RT distributions, thus allowing for a sophisticated investigation of the response process associated with faking. Future research can examine different parametrizations of the model and conduct validation analyses of particular model parameters.

References

Alagöz, Ö. E. C., & Meiser, T. (2024). Investigating heterogeneity in response strategies: A mixture multidimensional IRTree approach. *Educational and Psychological Measurement, 84*(5), 957–993. https://doi.org/10.1177/00131644231206765

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*(1), 29–51. https://doi.org/10.1007/bf02291411

Böckenholt, U. (2014). Modeling motivated misreports to sensitive survey questions. *Psychometrika, 79*(3), 515–537. https://doi.org/10.1007/s11336-013-9390-9

Bolt, D. M., & Meng, L. (2025). IRT-based response style models and related methodology: Review and commentary. *British Journal of Mathematical and Statistical Psychology*. Advance online publication. https://doi.org/10.1111/bmsp.70006

Brown, A., & Böckenholt, U. (2022). Intermittent faking of personality profiles in high-stakes assessments: A grade of membership analysis. *Psychological Methods, 27*(5), 895–916. https://doi.org/10.1037/met0000295

Ellingson, J. E., & McFarland, L. A. (2011). Understanding faking behavior through the lens of motivation: An application of VIE theory. *Human Performance, 24*(4), 322–337. https://doi.org/10.1080/08959285.2011.597477

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*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Dias, J. G., & Vermunt, J. K. (2008). A bootstrap-based aggregate classifier for model-based clustering. *Computational Statistics, 23*(4), 643–659. https://doi.org/10.1007/s00180-007-0103-7

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

Fine, S., & Pirak, M. (2016). Faking fast and slow: Within-person response time latencies for measuring faking in personnel testing. *Journal of Business and Psychology, 31*(1), 51–64. https://doi.org/10.1007/s10869-015-9398-5

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Griffith, R. L., & Converse, P. D. (2011). The rules of evidence and the prevalence of applicant faking. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 34–52). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0018

Hendy, N., Krammer, G., Schermer, J. A., & Biderman, M. D. (2021). Using bifactor models to identify faking on Big Five questionnaires. *International Journal of Selection and Assessment, 29*(1), 81–99. https://doi.org/10.1111/ijsa.12316

Holden, R. R. (1995). Response latency detection of fakers on personnel tests. *Canadian Journal of Behavioural Science, 27*(3), 343–355. https://doi.org/10.1037/0008-400x.27.3.343

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63*(2), 272–279. https://doi.org/10.1037/0022-3514.63.2.272

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin, 30*(2), 161–172. https://doi.org/10.1177/0146167203259930

Hu, J., & Connelly, B. S. (2021). Faking by actual applicants on personality tests: A meta-analysis of within-subjects studies. *International Journal of Selection and Assessment, 29*(3–4), 412–426. https://doi.org/10.1111/ijsa.12338

Hsu, L. M., Santelli, J., & Hsu, J. R. (1989). Faking detection validity and incremental validity of response latencies to MMPI Subtle and Obvious items. *Journal of Personality Assessment, 53*(2), 278–295. https://doi.org/10.1207/s15327752jpa5302_6

Jasra, A., Holmes, C. C., & Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science, 20*(1), 50–67. https://doi.org/10.1214/088342305000000016

Klehe, U.-C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance, 25*(4), 273–302. https://doi.org/10.1080/08959285.2012.703733

Kuncel, N. R., Goldberg, L. R., & Kiger, T. (2011). A plea for process in personality prevarication. *Human Performance, 24*(4), 373–378. https://doi.org/10.1080/08959285.2011.597476

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology, 62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13

Leng, C. H., Huang, H. Y., & Yao, G. (2020). A social desirability item response theory model: Retrieve-deceive-transfer. *Psychometrika, 85*(1), 56–74. https://doi.org/10.1007/s11336-019-09689-y

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*(4), 523–547. https://doi.org/10.1007/bf02294327

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/bf02296272

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633–644. https://doi.org/10.1509/jmkr.45.6.633

McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189x018002005

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*(2), 348–355. https://doi.org/10.1037/0021-9010.88.2.348

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

Nagy, G., & Ulitzsch, E. (2022). A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models. *Educational and Psychological Measurement, 82*(5), 845–879. https://doi.org/10.1177/00131644211045351

O'Hagan, A., Murphy, T. B., & Gormley, I. C. (2012). Computational aspects of fitting mixture models via the expectation–maximization algorithm. *Computational Statistics & Data Analysis, 56*(12), 3843–3864. https://doi.org/10.1016/j.csda.2012.05.011

Plummer, M. (2017). *JAGS: Just Another Gibbs Sampler* (version 4.3.2) [Computer software]. https://sourceforge.net/projects/mcmc-jags/

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News, 6*(1), 7–11. https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf

Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics, 41*(3), 300–325. https://doi.org/10.3102/1076998616636618

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin, 135*(2), 322–338. https://doi.org/10.1037/a0014996

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*(4), 489–509. https://doi.org/10.1007/s10869-007-9038-9

Röhner, J., Schütz, A., & Ziegler, M. (2025). Faking in self-report personality Scales: A qualitative analysis and taxonomy of the behaviors that constitute faking strategies. *International Journal of Selection and Assessment, 33*(1), e12513. https://doi.org/10.1111/ijsa.12513

Sackett, P. R., & Walmsley, P. T. (2014). Which personality attributes are most important in the workplace? *Perspectives on Psychological Science, 9*(5), 538–551. https://doi.org/10.1177/1745691614543972

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology, 78*(6), 966–974. https://doi.org/10.1037/0021-9010.78.6.966

Seitz, T., Alagöz, Ö. E. C., & Meiser, T. (2025). Disentangling qualitatively different faking strategies in high-stakes personality assessments: A mixture extension of the multidimensional nominal response model. *Educational and Psychological Measurement, 85*(6), 1237–1277. https://doi.org/10.1177/00131644251341843

Seitz, T., Spengler, M., & Meiser, T. (2025). "What if applicants fake their responses?": Modeling faking and response styles in high-stakes assessments using the multidimensional nominal response model. *Educational and Psychological Measurement, 85*(4), 747–782. https://doi.org/10.1177/00131644241307560

Seitz, T., Wetzel, E., Hilbig, B. E., & Meiser, T. (2024). Using the multidimensional nominal response model to model faking in questionnaire data: The importance of item desirability characteristics. *Behavior Research Methods, 56*(8), 8869–8896. https://doi.org/10.3758/s13428-024-02509-x

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*(4), 298–321. https://doi.org/10.1177/0146621605285517

Sudman, S., Bradburn, N. M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. Jossey-Bass.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408. https://doi.org/10.1007/bf02294363

Tett, R. P., & Simonet, D. V. (2011). Faking in personality assessment: A "multisaturation" perspective on faking as performance. *Human Performance, 24*(4), 302–321. https://doi.org/10.1080/08959285.2011.597472

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Uglanova, I., Nagy, G., & Ulitzsch, E. (2025, January 30). *A mixture IRT model for handling different types of careless respondents*. PsyArXiv. https://doi.org/10.31219/osf.io/tgys3

Ulitzsch, E., Nestler, S., Lüdtke, O., & Nagy, G. (2024). A screen-time-based mixture model for identifying and monitoring careless and insufficient effort responding in ecological momentary assessment data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000636

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika, 87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology, 73*(S1), 83–112. https://doi.org/10.1111/bmsp.12188

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology, 75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204. https://doi.org/10.3102/10769986031002181 van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer.

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology, 17*(7), 755–774. https://doi.org/10.1002/acp.914

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research, 11*, 3571–3594.

Wolodzko, T. (2023). *extraDistr: Additional univariate and multivariate distributions* (version 1.10.0) [Computer software]. https://cran.r-project.org/web/packages/extraDistr/index.html

Youngflesh, C. (2018). MCMCvis: Tools to visualize, manipulate, and summarize MCMC output. *Journal of Open Source Software, 3*(24), 640. https://doi.org/10.21105/joss.00640

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*(2), 168–190. https://doi.org/10.1177/1094428104263674

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What is the nature of faking? Modeling distinct response patterns and quantitative differences in faking at the same time. *Organizational Research Methods, 18*(4), 679–703. https://doi.org/10.1177/1094428115574518

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). Faking: Knowns, unknowns, and points of contention. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 3–16). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0011

Appendix

Pre-Study to Assess Response Time Behavior

As discussed in the Main Text, the literature is split concerning the overall effect of faking on response times (RT), with some studies finding shorter RTs (e.g., Holden et al., 1992; Holden, 1995; Hsu et al., 1989) and other studies finding longer RTs associated with faking (e.g., Fine & Pirak, 2016; Holtgraves, 2004; Walczyk et al., 2003). However, the studies differ in the operationalization of "faking". Whereas some used direct instructed-faking manipulations (e.g., Holden et al., 1992; Holden, 1995; Hsu et al., 1989), others used manipulations that should elevate desirability concerns in a more subtle way, for instance, by telling participants that their responses would be used to create a personality profile of them (Holtgraves, 2004).

To facilitate our understanding of RT effects associated with the three response strategies modeled in this article, and to empirically corroborate our assumption of increased RTs in the "S&F class", we conducted a pre-study using an experimental design. Therefore, we collected data from $N = 72$ participants (gender: 52.8% female, 47.2% male; age: $M = 22.83$ years, $SD = 3.18$, $range = [19, 35]$) on the online platform *SoSci Survey* (https://www.soscisurvey.de/) and instructed all of them to respond to a set of personality items in three conditions. In one condition (analogous to the "S-only class" in our model), participants should be fully honest in their responses, that is, they should base their responses on their true personality. In a second condition (analogous to the "S&F class"), participants should in principle also be honest but embellish their responses in the direction of social desirability. In a third condition (analogous to the "F-only class"), participants should base their responses solely on the items' desirability, without referring to their actual personality. The order of conditions was counterbalanced between participants. The items were the actual items of the test used in the high-stakes assessment from our empirical demonstration. Responses were given on a 7-point Likert scale (1 = *very low agreement* to 7 = *very high agreement*). Like in the actual assessment from our empirical demonstration, each item appeared on a separate questionnaire page, which allowed exact measurement of RTs on the item level. The order of items within conditions was also counterbalanced between participants. More information can be accessed at https://osf.io/crmv4/?view_only=db3eb1f7139742f48ee01887363b5e4a.

For analysis, we ran a series of multilevel regression models using the R packages *lme4* (Bates et al., 2015) and *lmerTest* (Kuznetsova et al., 2017), and set the significance level to $\alpha = .05$. We excluded RTs below 1 second and above 30 seconds, and log-transformed RTs to approximate a normal distribution. With individual observations nested in persons as well as items, we specified random intercepts for persons and random intercepts for items (i.e., cross-classified random effects). Table A1 shows the results of the fitted multilevel models. In the model without person and item covariates, the "S&F class" condition was associated with significantly longer RTs compared to the "S-only class" condition, whereas the "F-only class" was associated with significantly shorter RTs compared to the "S-only class" condition. These effects remained significant when adding person (gender, age) and item covariates (item keying, number of characters per item). Descriptively, the effects of the conditions were rather small, with median RTs of $Mdn = 3.87$ seconds ($MAD = 1.68$) in the "S-only class" condition, $Mdn = 4.32$ seconds ($MAD = 2.36$) in the "S&F class" condition, and $Mdn = 3.67$ seconds ($MAD = 1.79$) in the "F-only class" condition. A repeated-measures analysis of variance (ANOVA) yielded an effect size of $\eta^2_G = .051$ ($F(2, 142) = 19.69, p < .001$).

With these results, the pre-study provides evidence for existing RT differences between the different response strategies. In particular, RTs were longest when participants responded according to substantive traits *and* desirability, supporting the $\lambda > 0$ constraint in our proposed model. Also, the multilevel analyses yielded substantial random intercept variances for both persons and items. These were not considerably reduced when adding person and item covariates, emphasizing the importance of having a person speed parameter and item-specific time intensity parameters in the proposed model.

Table A1

Parameter Estimates, Standard Errors, and Model Fit Indices of the Multilevel Models of the Pre-Study

	Parameter estimate (standard error)				
Effect	Model 0	Model 1	Model 2	Model 3	Model 4
Fixed effects:					
(Intercept)	1.40 (0.04)***	1.38 (0.04)***	1.37 (0.05)***	1.34 (0.04)***	1.33 (0.05)***
Conditions:					
“S&F class” condition		0.12 (0.01)***	0.12 (0.01)***	0.12 (0.01)***	0.12 (0.01)***
“F-only class” condition		–0.04 (0.01)***	–0.04 (0.01)***	–0.04 (0.01)***	–0.04 (0.01)***
Person covariates:					
Gender			0.02 (0.06)		0.02 (0.06)
Age			0.05 (0.03)†		0.05 (0.03)†
Item covariates:					
Keying				0.10 (0.03)**	0.10 (0.03)**
Number of characters				0.01 (0.00) ***	0.01 (0.00) ***
Random effects:					
Level 1 residual variance	0.20	0.20	0.20	0.20	0.20
Level 2 residual variance due to persons	0.07	0.07	0.07	0.07	0.07
Level 2 residual variance due to items	0.02	0.02	0.02	0.01	0.01
Model fit indices:					
Log-likelihood	–6608.3	–6484.6	–6483.1	–6468.3	–6466.8
Deviance	13216.6	12969.2	12966.2	12936.6	12933.7
AIC	13224.6	12981.2	12982.2	12952.6	12953.7
BIC	13253.5	13024.5	13040.1	13010.4	13026.0

Note. $N = 72$ (10209 observations). Condition (“S-only class” condition as reference), gender (male as reference), and item keying (positively-keyed as reference) were dummy-coded; age was z-standardized; number of characters per item was centered. AIC = Akaike information criterion; BIC = Bayesian information criterion. p-values are two-tailed.

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.