Citation: Brown, A., & Bockenholt, U. (2021). Intermittent Faking of Personality Profiles in High-Stakes Assessments: A Grade of Membership Analysis. *Psychological Methods.*

Intermittent Faking of Personality Profiles in High-Stakes Assessments: A Grade of Membership Analysis

Anna Brown

School of Psychology, University of Kent

Ulf Böckenholt

Kellogg School of Management, Northwestern University

Author footnotes:

Correspondence should be addressed to Anna Brown, School of Psychology, University of Kent, Canterbury, Kent, CT2 7NP, United Kingdom. Tel: +44 (0)1227 823097. E-mail: a.a.brown@kent.ac.uk.

Acknowledgements:

Anna Brown thanks the University of Kent for granting her study leave to facilitate uninterrupted work on this research.

Abstract

In high stakes assessments of personality and similar attributes, test takers may engage in impression management (aka *faking*). This paper proposes to consider responses of every test taker as a potential mixture of '*real*' (or retrieved) answers to questions, and '*ideal*' answers intended to create a desired impression, with each type of response characterized by its own distribution and factor structure. Depending on the particular mix of response types in the test taker profile, *grades of membership* in the 'real' and 'ideal' profiles are defined. This approach overcomes the limitation of existing psychometric models that assume faking behavior to be consistent across test items. To estimate the proposed Faking-as-Grade-of-Membership (F-GoM) model, two-level factor mixture analysis is used, with two latent classes at the response (within) level, allowing grade of membership in 'real' and 'ideal' profiles, each underpinned by its own factor structure, at the person (between) level. For collected data, units of analysis can be item or scale scores, with the latter enabling analysis of questionnaires with many measured scales. The performance of the F-GoM model is evaluated in a simulation study, and compared against existing methods for statistical control of faking in an empirical application using archival recruitment data, which supported the validity of latent factors and classes assumed by the model using multiple control variables. The proposed approach is particularly useful for high-stakes assessment data and can be implemented with standard software packages.

Keywords: faking, impression management, social desirability, ideal-employee factor, Retrieve-Edit-Select, grade of membership

Intermittent Faking of Personality Profiles in High-Stakes Assessments: A Grade of Membership Analysis

Impression management (aka *faking*) is prevalent in assessments of personality, competencies and other self-reported attributes that carry important consequences for test takers, for instance, in high-stakes assessments. König and colleagues (2011) used a randomized response technique to ask test takers about past dishonest behaviors in job applications and found 32% of US study participants to have exaggerated their positive characteristics on questionnaires and 15% to have given completely false responses at some point in their lives. Robie, Brown and Beaty (2007) experimentally elicited an applicant context and observed participants as they were 'thinking aloud' how to respond. They found that at its extreme, faking behavior manifested itself in the production of responses that the participants thought typical of an 'ideal' applicant for the advertised job. However, they found a remarkable heterogeneity in the participants' behavior, with some simply recalling their attributes without any considerations of the 'ideal' applicant, and some bringing these considerations in to verify whether the recalled attributes were appropriate to report.

When for the same question some candidates report what they believe to be their real attribute and others report what they believe will increase their chances of being selected, *construct validity* is compromised by definition. When a test no longer measures what it is supposed to measure, we lack the foundation to meaningfully discuss such substantive issues in selection as, for instance, criterion-related validity (what are the common sources of variance that the test score and the criterion share?) or fairness considerations (do candidates with an equal true score on the target attribute but from different groups have equal chances of being selected?). Because construct validity plays a pivotal role in psychological assessment, detecting and controlling test faking is vital.

Despite significant progress in developing methods for statistical control of faking behavior, none have proven entirely satisfactory. Methods based on observed/manifest variables such as ‘social desirability’ or ‘lie’ scales (e.g. Paulhus, 1991) have been criticized for their inability to separate situational faking factors from substantive personality factors (McCrae & Costa, 1983). It is indeed doubtful that self-reporting of faking behaviors is helpful, since people may manipulate these scales in the same way as they manipulate other scales. Latent variable models are more promising since they incorporate unobserved effects underpinning the observed response behavior. These models can incorporate continuous latent *factors* and discrete latent *classes* into the response process. Individual differences in faking behavior are typically modeled by assigning to each person a score on a latent continuum, for instance an ‘ideal-employee factor’ (Klehe et al., 2012; Schmit & Ryan, 1993), or a membership in a latent class, for instance ‘honest responders’, ‘slight fakers’ and ‘extreme fakers’ (Zickar et al., 2004). In either case, the rather restrictive assumption is made that a person’s faking behavior is consistent across all questionnaire items.

There are several reasons to doubt the plausibility of this assumption. First, misreporting is unnecessary where the test taker has nothing to conceal (Tourangeau & Yan, 2007), thus the perceived need for faking responses may vary across scales or items. Second, such test taker goals as ‘staying true to self’, which have been identified by researchers as a powerful motive in high-stakes assessments, conflict with the more obvious motive of ‘appearing impressive’ (Kuncel et al., 2011), thus rendering the events of either all or none of the responses being faked as unlikely. Similarly, extant research on cheating has consistently shown that people will cheat if given an opportunity, but the vast majority will keep cheating to a minimum in order to maintain their self-image of an honorable person (Ariely, 2012). All this strongly suggests that many test takers will self-enhance on some but not all of the measured attributes.

To address this phenomenon of *intermittent* faking in high-stakes assessments, we suggest a comprehensive yet pragmatic method of modeling response behavior that considers any given person's profile as a potential mixture of true and faked scores. The proposed method adopts the view that for each presented question, test takers decide whether to report their actual standing or to report a standing that would look favorably on the job application (or whatever criteria the test takers are trying to fulfill). In this regard, our approach is based on research about misreporting in different contexts – high-stakes assessments, sensitive survey questions, strategic self-presentation and socially desirable responding (Böckenholt, 2014; Holtgraves, 2004; Kuncel et al., 2011; Robie et al., 2007; Tourangeau & Yan, 2007; Walczyk et al., 2005). However, the unique contribution of our approach is that we incorporate this notion into a psychometric model that can be readily applied to questionnaire data to identify the different person-level and item- or scale-level effects and, ultimately, enables a more nuanced insight into construct validity. The modeled effects include the substantive (true) scores on measured attributes, the respondents' propensity to edit their (true) scores in general and decisions to edit specific attributes, and the respondents' tendency to self-enhance once the decision to edit had been made.

Our presentation of the proposed approach is organized as follows. First we summarize existing approaches to control for faking behavior statistically and discuss their strengths and weaknesses. Then we present our psychometric model, which combines elements of the Retrieve-Edit-Select framework (Böckenholt, 2014) and grade-of-membership factor analysis (Asparouhov & Muthen, 2007), and discuss its estimation and parameter recovery. We then illustrate how attempts to control for intermittent faking when it is present with current methods that assume faking behavior to be consistent within a person can lead to biased conclusions. The practicality and usefulness of the proposed approach is demonstrated in an analysis of a large operational dataset. Here we show that personality profiles obtained in high-stakes assessments can be analyzed to infer the underlying true scores as well as the extent of intermittent faking; and that these inferences validate well against external variables. In the conclusion section, we discuss the strengths and limitations of the proposed approach and suggest some avenues for future research. To enable researchers to apply the proposed methodology, we supply sample code for Mplus (Muthén & Muthén, 2017) in an online Supplement.

Observed Effects of Faking Behavior

Empirical research provides ample evidence that high-stakes responses fail to comply with measurement models developed for low-stakes settings (e.g. Birkeland et al., 2006; Schmit & Ryan, 1993). Psychometrically, faking behavior can manifest itself in multiple ways:

1) Skewed distributions of item and scale scores. Distributions of scores for desirable/undesirable attributes are negatively/positively skewed and often show ceiling/floor effects, respectively (Birkeland et al., 2006). Distributions of scores for ambivalent attributes can even be bimodal (Kuncel & Borneman, 2007).

2) Inflated correlations of item scores, and of scale scores. All desirable attributes correlate positively (and negatively with undesirable attributes) even when they are conceptually unrelated, that is, when they should not systematically co-occur in individuals (Klehe et al., 2012; Schmit & Ryan, 1993).

3) Inflated estimates of internal consistency reliability (alpha). This is an artifact of the inflated correlations of item scores.

A Review of Current Modeling Approaches

To model the observed effects of faking behavior, various approaches involving latent variables – continuous and categorical – have been proposed. To generalize the presentation of these response models to different types of observed variables – continuous, ordinal or binary – we consider latent *response tendencies* y^*. When the observed responses y are categorical, logit or probit link functions can be used to model the response tendencies (McDonald, 1999). When the observed responses y are continuous, the identity link is used.

We apply a linear factor analysis (LFA) measurement model, in which response tendencies are underlain by linear combinations of d common factors $\boldsymbol{\theta}^{(\mathrm{R})}$ as illustrated in Figure 1a. With this, person j has the following tendency for response i:

$$y_{ij}^* = \mu_i + \sum_{q=1}^{d} \lambda_i^q \theta_j^{(\mathrm{R}q)} + \varepsilon_{ij}\,, \qquad (1)$$

where μ_i is the mean response, $\boldsymbol{\theta}_j^{(\mathrm{R})} = \left(\theta_j^{(\mathrm{R1})}, \theta_j^{(\mathrm{R2})} ..., \theta_j^{(\mathrm{R}d)}\right)'$ is the vector of person scores on d measured attributes (common factors), λ_i^q is the factor loading on attribute q, and ε_{ij} are random realizations of the unique factor. We add the superscript (R) to highlight that the measured attributes are the '**R**etrieved' attributes (Böckenholt, 2014) – we will elaborate on the meaning of this later – but for now, it suffices to think of them as 'substantive' attributes the questionnaire is designed to measure. The common factors $\boldsymbol{\theta}^{(\mathrm{R})}$ are assumed multivariate standard normal. The unique factors are assumed standard normal (or standard logistic if the logit link is used for categorical responses) and uncorrelated with all the common factors as well as with each other. Typically, questionnaires are designed so that for any given item all but one loading are zero, forming subsets of items measuring just one attribute in common while collectively all items measure multiple attributes – conforming to the so-called "independent clusters" basis (McDonald, 1999).

INSERT FIGURE 1 ABOUT HERE

Common Method or 'Ideal-Employee' Factor (IEF)

As illustrated in Figure 1b, the effect of stakes on item responses can be modeled via a general factor affecting all items (accounting for the 'method' variance) in addition to factors specific to conceptually concordant item sets (accounting for the 'substantive' variance due to individual differences on the measured attributes). The 'common method' factor model assumes the following tendency for response *i* of person *j*:

$$y_{ij}^{*} = \mu_i + \sum_{q=1}^{d} \lambda_i^q \theta_j^{(\mathrm{R}q)} + s_i f_j + \varepsilon_{ij}\,. \tag{2}$$

Equation (2) differs from the basic Equation (1) by the addition of factor f_j common to **all** items. The loading s$_i$ reflects the extent of over-reporting on *i* (or under-reporting if s$_i$ is negative) for someone with the score f_j = 1. Factor *f* has been established in factor analyses of applicant data (Schmit & Ryan, 1993); it has been named '*ideal-employee factor*' (IEF) because it yields a pattern of responses deemed typical of an 'ideal employee' by applicants with a high overall tendency to over-report. The distribution of *f* is assumed standard normal, and for identification purposes, it is set orthogonal to the substantive factors. For questionnaires designed on the "independent clusters" basis, each response would be influenced by just two common factors – method-related and substantive – thus reducing Equation (2) to a type of bi-factor model with correlated specific factors.

Strengths and weaknesses of the IEF approach. The IEF model is a popular choice for the analysis of large questionnaires. When faking effects are notable, modeling the IEF has been shown to: (1) improve goodness of fit for measurement models; (2) overcome the inflation of correlations between measured scales; and (3) increase convergent and reduce discriminant correlations with external measures (e.g. A. Brown et al., 2017; Schmit & Ryan, 1993).

Despite these advantages, the IEF model has important deficiencies. First, the constraint of orthogonality of IEF to all the substantive factors may be unreasonable; for example, people may be more or less likely to over-report when their true scores are low. Despite this orthogonality constraint, the model is often empirically under-identified (Podsakoff et al., 2003), which is also a problem. The measurement of only one substantive attribute is one obvious situation in which the IEF model is under-identified, with Equation (2) reducing to an exploratory two-factor model. Second, the IEF model does not explain skewed distributions observed in high-stakes assessments. When all effects in Equation (2) are specified to be symmetrically (normally or logistically) distributed, response tendencies and substantive factor scores estimated from them must also be symmetrically distributed but, as noted previously, in practice they are not. Moreover, the additive effects in the IEF model conceptualize faking as an "over-report" bias that test takers add to their retrieved scores. This notion is inconsistent with the discrete nature of faking decisions identified in qualitative and experimental research. Robie and colleagues (2007) show that when respondents activate their real attributes, they may not think from the perspective of an 'ideal employee'; and when they activate the ideal response, they may not think about where they stand on the attribute in question. Even when these two considerations take place at the same time, whether the test taker chooses to give the true or desirable response is a discrete event (Leite & Cooper, 2010).

Factor Mixture Analysis (FMA)

To address the heterogeneity in distributions of observed scores, it can be assumed that test takers come from two or more unobserved subpopulations, each utilizing different response processes. The subpopulations are operationalized through a categorical latent (class) variable with two or more levels, for instance capturing applicants who report their actual attributes and who report attributes deemed typical for an ‘ideal employee’. A factor model may be specified to account for within-class dependencies among responses indicating the same attribute; however, the parameters of this measurement model may differ between classes, as illustrated in Figure 1c.

Factor mixture analysis (FMA) provides a general framework appropriate for such modeling (Clark et al., 2013; Muthen, 2007) by describing the tendency for response *i* as conditional on person *j*’s membership in class *c*:

$$y_{ij}^{*}\left(C_{j}=c\right)=\mu_{ic}+\sum_{q=1}^{d}\lambda_{ic}^{q}\theta_{j}^{q}+\varepsilon_{ijc}\,. \tag{3}$$

Equation (3) adopts the notation of earlier equations, and subscript *c* signifies class-specific measurement parameters. Note that we no longer have superscript (R) for the common factors, because the most general formulation of FMA allows for different factors in each class defined by class-specific loadings on a common set of factors. For example, we may want to model substantive (‘Retrieved’, denoted ‘R’) factors in the ‘real’ class, but a different structure underpinning self-presentation in the ‘ideal’ class – perhaps a structure with fewer factors. This can be accommodated by fixing some of the factor loadings in Equation (3) to zero in some classes.

Particular contexts may call for more restricted versions of the FMA model. For instance, Zickar et al. (2004) investigated class differences in thresholds for ordinal items measuring a single factor, assuming equal item discriminations (polytomous extension of the Rasch model). Three classes were modelled – ‘honest’, ‘slight faking’ and ‘extreme faking’ – each assuming a single factor with all loadings fixed to 1 in all classes, thus fixing the factor meaning and allowing variation in threshold behavior only. Less restricted FMA versions will be used in the present paper. In our empirical Application, we will model two classes, with five factors in the ’honest’ class (*d* =5 in Equation (4)), and just one factor in the ’faking’ class. In this case, the meaning of factors will be defined according to theory presented later and denoted ‘R’ for ‘Retrieved’ or substantive factors, and ‘S’ for ‘Selected’ or presentational factor.

$$\left[y_{ij}^{*} \middle| C_{j}\right] = \begin{cases} \mu_{i0} + \sum_{q=1}^{d} \lambda_{i}^{q} \theta_{j}^{(\mathrm{R}q)} + \varepsilon_{ij0}, & \text{if } C_{ij} = 0 \\ \mu_{i1} + s_{i} \theta_{j}^{(\mathrm{S})} + \varepsilon_{ij1}, & \text{if } C_{ij} = 1 \end{cases} \tag{4}$$

Strengths and weaknesses of the FMA approach. FMA models incorporating a small number of classes are relatively easy to fit. The key advantage of FMA models is that they can explain skewed distributions and heteroscedasticity of relationships with other variables observed in applicant data (Pavlov et al., 2019) by conceptualizing the population as a mixture of two or more normally distributed subpopulations. FMA also accounts for the high inter-correlations observed in applicant data, by allowing for different item means in different subpopulations of applicants. This ‘main effect’ of the latent classes plays a similar role to the ideal-employee factor, but with a finite set of values this effect is allowed to take.

However, the FMA approach has important deficiencies too. First, the class membership is fixed, which implies that test takers are assumed to either respond honestly to **all** items, or to fake **all** of them. In this respect, FMA is similar to IEF, which also assumes consistent response behavior, given a person’s tendency to over-report. However, as we argued in the introduction, this specification may prove to be too restrictive as test takers may fake responses on some items or scales but provide honest responses on other items or scales. Unable to account for this within-person heterogeneity, FMA often yields intermediate classes, for example ‘slight fakers’ (Zickar et al., 2004). We believe such classes to be methodological artefacts, as it seems unlikely that some test takers will intentionally produce test profiles that are only ‘half-good’ to get a job.

‘Retrieve-Edit-Select’ (R-E-S) Decision Model

To account for the item-by-item basis for discrete decisions to either give one’s true response or to conceal it, Böckenholt (2014) suggested employing as many binary latent class variables as there are items. Within each item, assumed to have several ordered response options, a decision process with three stages is described. At the first ‘Retrieve’ stage, test takers retrieve an answer, presumably by recalling their typical behavior, attitude, etc. The retrieved answer is underlain by $\theta^{(R)}$ – a substantive attribute that the item is designed to measure, via a probit link function with equal discrimination parameters. At the second ‘Edit’ stage, respondents decide whether to report the retrieved answer or to edit it. This binary decision is influenced by item-specific considerations but also the overall propensity to edit one’s responses $\theta^{(E)}$ (superscript (E) marks the ‘Edit’ stage), again via a probit link function with equal discriminations. If the decision to edit has been made, the third ‘Select’ stage is engaged. In sensitive survey questions, a response category is selected that suitably conceals the retrieved answer. This process is realized via transition functions that describe probabilities of selecting higher categories (or lower, but this direction is assumed consistent throughout the survey) as conditional on the retrieved response category. The transition is driven by the individual preference $\theta^{(S)}$ (superscript (S) marks the ‘Select’ stage) for extreme response categories, via a logit link similar to the IRT partial credit model.

The RES model is schematically presented in Figure 1d, which features three latent factors: the substantive attribute $\theta^{(R)}$ underlying retrieved answers, the tendency to edit responses $\theta^{(E)}$, and the tendency to select extreme response categories $\theta^{(S)}$. Interactions between retrieved responses r_i and $\theta^{(S)}$ – the transition functions – are presented as filled circles.

Strengths and weaknesses of the R-E-S approach. The R-E-S model is theoretically appealing because it conceptualizes faking as an editing process at the item level, which may or may not result in adjustments but always increases the response time, as experimental research has shown (Holtgraves, 2004; Walczyk et al., 2005). Another point of strength is the confirmatory nature of the R-E-S model because response determinants at every stage of the process are specified fully. This dictates clear identification constraints, such as the pre-defined and fixed direction of editing (for example, people may over-report knowledge but do not under-report it). Another advantage compared to FMA, which assumes mutually exclusive classes and thus estimates different effects for individuals in 'honest' and 'faking' classes, the R-E-S model allows exploring the relationship between constructs underpinning decisions of both types by modeling them for the same individual.

However, modeling as many latent class variables as there are items (with the overall number of classes for the response pattern of length *m* equal 2^m) and three latent tendencies comes at a price – the R-E-S model is demanding to estimate. More importantly, the number of Retrieve factors is limited to one, and its effect on all responses is assumed uniform, which is limiting in personality research. The R-E-S model was developed for a specific context – responding to sensitive survey questions, and in particular, may not adequately describe the Select stage when it comes to personality measures. In sensitive survey questions, recognizing the most or the least revealing response option is trivial, thus respondents differ only in the degree of appeal they feel for the extreme categories, calling for a uniform effect of $\theta^{(S)}$ on the items (Böckenholt, 2014). In personality assessments, the most desirable response option is often not trivial and depends on the item content and testing context. For example, strongly agreeing with an item describing assertive behavior might be advantageous when applying for a sales role; but it might be detrimental for a supporting role, where such behaviors are less desirable. More generally, many test takers believe that endorsing the extreme response option for some questions may not be believable or may come across as arrogant (Kuncel & Tellegen, 2009). To reflect these realities of the personality assessment context, the Retrieve and Select stages need a more general formulation. Specifically, the respondent's perception of the importance of a presented attribute to the selection criteria, as well as the most desirable level of the attribute to fulfill these criteria play a central role in deciding what answer to pick, thus necessitating item-specific effects.

Proposed Model: Faking as Grade of Membership in 'Real' and 'Ideal' Profiles

The existing latent variable models solve some of the problems posed by faking distortions; however, none solve them completely. The Retrieve-Edit-Select framework of Böckenholt (2014) offers the most comprehensive foundation for describing faking behavior in high-stakes assessment contexts. The Edit stage of the RES framework can be adopted directly; however, more general models are needed for the Retrieve and Select stages. Moreover, such models need to be formulated to allow for the estimation of realistic assessments involving several measured attributes and many items. Next, we present an approach that addresses these objectives.

We assume that when responding to questionnaires, test takers activate or **retrieve** information relevant to attributes being measured (Robie et al., 2007). If the stakes are low, and test takers are not motivated to manage impression of selves, the retrieved answer is reported. We will refer to such answers as 'real', reflecting the perceived reality of this information by test takers, although it is acknowledged that it could be subject to unconscious biases. If the stakes are high, and what is reported has important consequences, test takers may engage in self-censoring, whereby they appraise the retrieved answer and decide whether it needs to be **edited** (Holtgraves, 2004). Test takers may decide not to edit and report the retrieved answer. Or, they may decide to edit, and **select** a response that would create a desired impression and maximize their chances to be selected.

Responses at the Retrieve and Select stages have different antecedents. While the retrieved responses reflect the levels of personal attributes the test taker possesses (i.e., what the questionnaire is designed to measure), the selected responses reflect the levels of attributes that, in the test taker's opinion, would impress the employer (Kuncel et al., 2011; Kuncel & Borneman, 2007). Individual differences around the average opinion about suitable levels of selected responses may be caused in part by personal tendencies to produce extreme scores. Studies in which people are explicitly instructed to produce an ideal profile for a specific job or criteria are useful in understanding the features of these Select-stage responses. Although people's views of the ideal profile may differ, particularly for ambiguous attributes (Kuncel & Borneman, 2007), it has been found that 'ideal' responses show less variation than retrieved responses and simpler factor structures (Haaland & Christiansen, 1998; Zickar et al., 2004). These distributional differences can be used to specify the archetypal retrieved and selected profiles, each with its corresponding means and covariance structure.

Thus we propose to model two basic profile types – 'real' or retrieved (R) and 'ideal' or selected (S) – with responses indicating one or the other profile when person *j* decides not to edit ($C_{ij} = 0$) or edit ($C_{ij} = 1$) response *i*, respectively:

$$\left[y_{ij}^{*} \middle| C_{ij} \right] = \begin{cases} \mu_{i0} + \sum_{q=1}^{d} \lambda_i^q \theta_j^{(\mathrm{R}q)} + \varepsilon_{ij0}, & \text{if } C_{ij} = 0 \\ \mu_{i1} + s_i \theta_j^{(\mathrm{S})} + \varepsilon_{ij1}, & \text{if } C_{ij} = 1 \end{cases} \tag{5}$$

While we allow several substantive factors[1] $\theta_j^{(\mathrm{R}1)}, \ldots, \theta_j^{(\mathrm{R}d)}$ to underlie the retrieved responses, we propose that only one factor $\theta_j^{(\mathrm{S})}$ underlies the selected responses. This factor

[1] However, it may be computationally challenging to estimate more than a few retrieved factors with maximum likelihood estimators, as explained in the 'Implementation in a Multilevel Framework' section.

captures individual differences in the extremity of selected scores compared to the population mean; there will be those maximizing scores on desirable items and minimizing on undesirable much more than the average candidate, producing a rather extreme ‘ideal’ profile, and there will be those who produce a much more understated ‘ideal’ profile than the average candidate.

Beyond the prescribed factor structures for the two classes in Equation (5) versus the open structures in Equation (3), the model described above differs from the FMA model in one very important aspect. That is, the subscript *i* is added to class membership C_{ij}, signifying that the decision to edit is made **response by response**, and therefore editing some responses and not editing others permits memberships in both ‘real’ and ‘ideal’ profiles. One important consequence of the same person engaging in both types of responding is that the substantive and presentational effects, $\boldsymbol{\theta}^{(R)}$ and $\theta^{(S)}$, are modeled for the same person, thus enabling exploration of the relationship between them. This is in contrast to the FMA model, where the profile membership is mutually exclusive, and the relationship between $\boldsymbol{\theta}^{(R)}$ and $\theta^{(S)}$ cannot be established.

To describe behavior at the Edit stage, we describe the probability of editing response *i* as the logistic function of the person *j*’s overall tendency to edit responses $\theta_j^{(E)}$ coupled with an item-specific intercept α_i:

$$P\left(C_{ij}=1\right)=\frac{\exp(\alpha_i+\theta_j^{(E)})}{1+\exp(\alpha_i+\theta_j^{(E)})}. \tag{6}$$

Beyond the overall propensity to edit one’s responses $\theta^{(E)}$, a normally distributed variable with the mean zero and variance estimated from the data, the decisions to edit depend on the intercept α_i, controlling for the average rate of editing response *i*. This rate is likely dependent on whether the item represents an important assessment criterion. For instance, if test takers see stress tolerance as a critical job requirement, they are likely to edit retrieved responses on the item ‘I get stressed out easily’ more often than if this behavior was not relevant to the job. The hypothesis that the intercept α_i relates to the *importance* of the selection criterion can be tested by modeling α_i as a linear combination of an overall intercept α and the item’s importance x_i (for instance, expert-rated) weighted by a coefficient β common to all items:

$$\alpha_i = \alpha + \beta x_i . \tag{7}$$

Further possibilities for modelling intercepts α_i may involve person-specific covariates, such as the test taker’s judgements about the importance of attributes as selection criteria. A potentially important construct that affects these judgements is the ‘ability to identify [selection] criteria’ (ATIC), which represents a person’s propensity to detect which criteria are being assessed for instance in assessment centers (e.g., Klehe et al., 2012).

We note that the model presented here is a factor mixture analysis extended to allow for partial class membership, the so-called ‘Grade of Membership’ (GoM) extension of FMA described by Asparouhov and Muthén (2007). We therefore refer to the proposed model as ‘*Faking as Grade of Membership*’ or *F-GoM* model. As also shown in an empirical Application presented further in the paper, allowing for partial membership in the ‘real’ and ‘ideal’ profiles is crucial in addressing the item-specific response decisions taking place in high stakes assessments. Boundary conditions include contexts in which the decision to edit (or not) is made a priori, and people edit none or all of the retrieved responses, demonstrating pure ‘real’ and ‘ideal’ profiles respectively. The former is probably the case in research settings, and the latter might be the case in instructed faking studies. Such cases can be accommodated by the F-GoM model by letting the variance of the tendency to edit $\theta^{(E)}$ approach infinity, in which case the model reduces to the respective two-class FMA model (Asparouhov & Muthen, 2007).

Implementation in a Multilevel Framework

To facilitate the application and use of the F-GoM model, in this section we discuss how to re-express this model so it can be estimated with standard software programs for multilevel mixture models. Asparouhov and Muthen (2007) point out that any GoM model with a factor structure is a special case of the general two-level mixture model, where the person *j* is a cluster, and the observed variables y_{ij} are a vector of univariate observations clustered in person *j*. Then, all item effects including the class membership C_{ij} are modeled at the within-person level, and all person effects including $\theta_j^{(R)}$, $\theta_j^{(S)}$ and $\theta_j^{(E)}$ are modeled at the between-person level. The only complication is that the F-GoM model requires that all measurement parameters (intercepts, factor loadings and residuals) vary from item to item, which is not feasible under a univariate treatment of the observed variables in the standard two-level mixture setup. To enable the required model features, we present each observed variable y_{ij} coming from record *i* within cluster/person *j* as L new dummy variables y_q, where L is the number of items, in such a way that:

$$y_{qij} = \begin{cases} y_{ij}, \text{ if } q = i \\ \text{missing, if } q \neq i \end{cases} . \qquad (8)$$

After re-arranging the data from the typical ‘wide’ format (one record per person, with L observed variables) to the ‘long’ format (L records per person), each observation now consists of L dummy variables, only one of which is not missing and corresponds to the observed value for that item and that person. Optionally, each record can also contain item-level covariates **x**, for example, the expert-rated importance of the item to the target job as per Equation (7).

Once the variables have been appropriately coded, the F-GoM model can be easily implemented using Mplus software (Muthén & Muthén, 2017). In a Supplement to this article, we provide example code for Mplus, including the suggested coding for observed variables. The observed variables y_{qij} exist at the within level which is also the level at which the class-varying residual variances var(ε_{ic}) are estimated. The class-varying means μ_{ic} and factor loadings λ_{ic} and s_{ic} are estimated at the between level, where the respective factors $\theta^{(R)}$ and $\theta^{(S)}$ are modeled. The optional observed item covariates, such as x_i described in Equation (7), exist at the within level only, where their effects on the item-level editing decisions are estimated. Finally, the person-level tendency to edit responses (to make item-level editing decisions C_{ij}=1) is also modeled at the between level. This tendency is denoted by c#1 in Mplus with reference to the first latent class, and is equal to $\theta^{(E)}$ in Equation (6) if the first class is the 'ideal' class; and it is equal $-\theta^{(E)}$ if the first class is the 'real' class.

Estimation

Maximum likelihood with robust standard errors (denoted MLR) is the default estimator for two-level mixture models in Mplus. The estimation requires as many dimensions of integration as there are person-level effects. Thus, at minimum the model requires three dimensions of integration – one for the Retrieve stage $\theta^{(R)}$, one for the Select stage $\theta^{(S)}$ and one for the Edit stage $\theta^{(E)}$. More dimensions may be required if the Retrieve stage involves multiple factors. With more than four dimensions, Monte Carlo integration with random points can be used. Useful starting values are provided by the estimates of the corresponding FMA model (Asparouhov & Muthen, 2007).

Recovery of model parameters: A Simulation Study

We conducted a simulation study to assess the performance of the F-GoM model in realistic conditions. This study examined the accuracy of F-GoM model parameter recovery when two factors are assessed at the Retrieve stage with continuous indicators that could be sum scores or any other observed measurements. The study design was aimed to replicate conditions in the Application presented later, namely various levels of correlation between latent factors at Retrieve, Edit and Select stages, the prevalence of editing, and the features of factor loadings at the Select stage. We tested the parameter recovery for two sample sizes: N=500 and N=1000. Mplus input code for the study is available as an online Supplement.

Data Generation. For this study, we assumed an assessment designed to measure two factors, with five continuous attributes (which could be scales scores) indicating each factor (10 attributes in total). We assumed that all attributes are positive and equally discriminating indicators of their respective substantive factors at the Retrieve stage (their factor loadings on $\theta^{(R1)}$ and $\theta^{(R2)}$ were set to 1). However, Attributes 1, 2 and 3 in each factor were assumed desirable for the target job, Attribute 4 ambivalent, and Attribute 5 undesirable. For example, Openness is indicated positively by such constructs as creativity (desirable), unconventionality (ambivalent desirability) and seeking variety (often undesirable). The desirable attributes yield higher means in selected than in retrieved responses, the undesirable attributes yield lower means, and the ambivalent attributes yield similar means. The Select-stage loadings were set to reflect the extent of over-reporting on attributes compared to the average ‘ideal’ profile. We note that desirable Attributes 2 and 3, elevated equally on average in ‘ideal’ profiles, had different loadings on the tendency to produce extreme profiles $\theta^{(S)}$. The magnitudes of factor loadings and residual variances at the Select stage were assumed lower than at the Retrieve stage, because usually, there is far more agreement (less variance) among people on what is ‘ideal’ than on what they report in their ‘real’ profiles (Haaland & Christiansen, 1998). All generating item parameters are reported in Table 1.

The substantive factors $\theta^{(R1)}$ and $\theta^{(R2)}$ were specified to be moderately correlated at .3. The tendencies to edit $\theta^{(E)}$ and to produce extreme profiles when editing $\theta^{(S)}$ were assumed uncorrelated. One substantive factor $\theta^{(R1)}$ was set positively correlated with both $\theta^{(S)}$ and $\theta^{(E)}$ at .5, and the other factor $\theta^{(R1)}$ was set to be uncorrelated with them. All structural parameters are reported in Table 2.

INSERT TABLES 1 AND 2 ABOUT HERE

Ten continuous 'Retrieved' values indicating $\theta^{(R1)}$ and $\theta^{(R2)}$, 10 continuous 'Selected' values indicating $\theta^{(S)}$, and 10 binary Edit stage decisions indicating $\theta^{(E)}$ were generated. The Edit stage decisions (either 0 or 1) were generated as binomial draws from a 1-parameter logistic IRT model in Equation (6), assuming that decisions to edit were taken half the time (all intercepts α_i set equal 0). Five hundred replications were generated with sample sizes of N = 1,000 and N = 500 each. From these generated values, 10 'observed' response variables Y_1-Y_{10} were derived as follows. For each person *j* and each scale *i*, the 'observed' response was set equal the Retrieved value if the corresponding decision was **not** to Edit, and it was set equal the Selected value otherwise.

Results. Table 1 summarizes the results pertaining to item parameters; Table 2 summarizes the results for structural parameters. For each parameter, we report the mean estimate across 500 replications, the mean standard error, and the 95% coverage (proportion of replications for which the 95% confidence interval contains the population parameter value).

For N = 1000, the estimated bias is very small (never exceeds 2.5%), the mean standard errors agree well with the standard deviations of estimated parameters, and the coverage is close to the correct 95%. Estimation accuracy for item parameters does not appear to depend on whether the substantive factor correlates with the faking tendencies or not. For N = 500, convergence was slower than for the larger sample, and three replications out of 500 did not converge. The estimated bias is larger than for N = 1000 but still small (never exceeds 4.5%), the mean standard errors are also larger, but they agree well with the standard deviations of the estimated parameters, resulting in a close to 95% coverage. We conclude that overall, the F-GoM model parameters and their SEs are recovered well.

Controlling for Intermittent Faking with Current Models

In this section, we discuss consequences of controlling for intermittent faking with models that assume consistent faking (FMA and IEF models discussed earlier). Simulated data provide useful insights here. To illustrate, we analyzed the 'observed' responses (which are mixtures of retrieved and selected responses) generated in the simulation study reported earlier not just with the F-GoM model but also with appropriate IEF and FMA models.

First, we illustrate how different models attribute the fact of editing to latent classes or factors designed to control for faking. Figure 2 shows scatterplots of two observed variables Y1 and Y2 generated according to the F-GoM model with parameters given in Tables 1 and 2. These two variables were indicators of the same retrieved factor $\theta^{(R)}$, and both had equal positive loadings on $\theta^{(S)}$. In Panel A of Figure 2, each bi-variate point (corresponding to a simulated individual) is marked by its true membership in one of four possible classes: (1) both Y1 and Y2 are 'real' responses; (2) both Y1 and Y2 are 'ideal' responses; (3) Y1 response is 'real' and Y2 response is 'ideal'; and (4) Y1 response is 'ideal' and Y2 response is 'real'. As both variables were generated as desirable, with 'ideal' means higher than 'real' means, the 'ideal-ideal' class is clearly clustered in the top right quadrant. The 'real-real' class spreads across the whole scatterplot centering on zero, and the two intermediate classes located in the left top and the bottom right quadrants. In Panel B of Figure 2, each point is marked according to its item-specific memberships in either 'real' or 'ideal' class as estimated by the F-GoM model. Despite occasional mistakes in classification of responses in overlapping clusters, it is clear that the model is capable of identifying intermittent faking.

In Panel C of Figure 2, the points are marked according to exclusive membership in either the 'honest' or 'faking' class as assigned by the FMA model with 2 classes. The model tends to assign correctly the 'real-real' responses to the 'honest' class and the 'ideal-ideal' responses to the 'faking' class; however, its assignments of inconsistently edited items (intermittent faking) are disproportionately influenced by the mean elevation from ‘real’ to ‘ideal’ responses for each variable. For instance, response on an item with larger ‘real-ideal’ mean difference (Y1) tends to have a bigger influence on membership assignment.

Finally, In Panel D of Figure 2, the same points are marked by their standing on the IEF factor score. For ease of illustration, we split this continuous score into four bands: low ($Z < -1$), below average ($-1 \le Z < 0$), above average ($0 \le Z < 1$) and high ($Z \ge 1$). It can be seen that the IEF banding sliced the scatter into clusters with similar sums of Y1 and Y2 – low, medium-low, medium-high and high, thus restricting the range of values in each cluster. Despite positive correlations between Y1 and Y2 within all ‘true’ classes (see the legend in Figure 2, panel A), correlations for each level of IEF are near-zero or negative (see the legend in panel D). The IEF model thus tends to over-extract the substantive variance shared by Y1 and Y2 and assign it to the method factor. We note that the FMA model does not have this tendency (see panel B).

INSERT FIGURE 2 ABOUT HERE

Finally, we consider the assessment of goodness of fit for non-mixture models (i.e. IEF models) when intermittent faking is present. When fitting an appropriate IEF model to the 500 datasets of N=1,000 each, generated as part of the simulation study reported earlier, we found that the chi-square statistic was insensitive to systematic misfit at the individual level. Thus, only a slightly larger proportion of replications (8.9%) than the nominal 5% level were rejected by the chi-square test, which is based on discrepancies between the observed and predicted covariance matrices, while discrepancies between observed and predicted individual values were large. This is not a new finding. Asparouhov and Muthén (2017) demonstrated that the usual fit statistics such as the chi-square test, CFI and TLI are not sensitive to misspecifications emerging from the failure to model an underlying heterogeneity in the population. To illustrate, Figure 3 depicts scatterplots of simulated responses Y1 from one of the replications against responses estimated by each of the alternative models we tested (the situation is similar across all replications). It can be seen that the IEF and FMA models yielded similarly poor accuracy of approximation of individual level values compared to the generating F-GoM model; however, the chi-square statistic (χ^2 (24, N = 500) = 22.6, *p* = 0.54) wrongly suggested that the IEF model was suitable. We therefore suggest that in future analyses of distortions caused by faking, researchers do not rely exclusively on covariance-based fit statistics but also study the discrepancies between the observed and predicted observations at the individual level, for instance using residual plots (Asparouhov & Muthén, 2017).

In contrast, the performance of information criteria AIC and BIC that are based on the individual-level information was satisfactory, as in the simulation studies they favored the correct model every time and by a very large margin. The AIC discriminated between the alternative models particularly well, and also correctly favored the IEF model over the CFA model with no control of faking at all, while the BIC criterion with its severe penalty on the number of estimated parameters favored the CFA model in some cases.

INSERT FIGURE 3 ABOUT HERE

Modeling Profiles at the Scale Level

Because faking behavior is an item-level phenomenon, analysis at the item level is an obvious default option. For categorical item responses, logit or probit link functions may be applied to the continuous response tendencies described in Equation (5). The downside of this approach is that it may be difficult to carry out in practice. One common challenge is the use of large questionnaires with many measured scales. Another is the unavailability of item-level responses, for instance because of the lack of access to test provider’s databases, or because of the loss of such access by test user organizations, or because only scores used in decision making for reporting purposes have been kept.

Because of these limitations and also because in practice personality profiles are reported and interpreted typically at the scale level, we need to consider the possibility of using scale scores as units of analysis (‘responses’) in the F-GoM model. Here the main analytic goal is to determine whether the reported scale score is likely to be from the ‘real’ or the ‘ideal’ profile; or whether the scale score largely represents the Retrieved standing on the scale or the Selected standing. Such analysis is easily done within the proposed F-GoM framework assuming continuous outcomes and the identity link applied to y^* in Equation (5). We note that we do not assume any particular mechanism for aggregation of item-level decisions to the scale level; rather, we suggest to think of scale scores as proxies for test takers’ self-reported standings on attributes the test is designed to measure. These descriptions, just like item responses, can be ‘real’ or ‘ideal’ and therefore can be analyzed for intermittent faking (in this case, ‘intermittent’ refers to faking some scales but not others). Clearly, the same approach can be applied to actual single-item measures, which sometimes comprise detailed definitions of behaviors or attitudes typical for persons with low, medium and high scores, for example via the dynamic analog scale (E. A. Brown & Grice, 2011).

We demonstrate in the Application section that modeling whole personality profiles consisting of over 30 scale scores is straightforward with F-GoM. The scales measured there are narrowly defined personality attributes, many of which show highly skewed distributions, rendering mixtures of ‘real’ and ‘ideal’ scores likely. We turn to the Application next.

Application: Analysis of Personality Profiles in Recruitment Settings

In this section, we illustrate the use of the proposed approach for analyzing operational recruitment data. As is often the case in operational settings, only scale-level assessments scores were kept in an archive and the item responses are no longer available. We show that despite the item aggregation into scales, it is possible to measure faking behavior and to correctly identify variance pertaining to the Retrieve, Edit, and Select stages, which can be validated by auxiliary variables collected during the assessments. Since the present application does not allow experimental control over antecedents of faking behavior, we adopt a validation strategy that tests a priori plausible relationships between the constructs identified by the faking model and independently measured variables. In the following, we first review the sample and available measures before we present the analysis and results.

Sample

We consider an archival dataset from a large US retail company, dating back approximately 10 years. The archive included data on applicants to operational and junior supervisory roles in different job families, with applications within each family assessed against different competencies. Specifically, we consider N =762 applications to ‘analytical’ job family roles, such as financial and process analysts, accountants, tax and logistics specialists. Approximately 49.3% of the applicants were men. The largest ethnic groups were White (76.2%), Asian (6.6%), and African American (5.2%), other groups accounting for less than 1%, and 5.6% of the applicants did not provide ethnicity information. The age of the applicants was not available to us.

Measures

Personality. The normative version of the Occupational Personality Questionnaire or OPQ, OPQ32n (Bartram et al., 2006) published by SHL, was used to measure 32 work-related personal styles grouped into four domains – Relationships with People, Thinking Styles, Feeling and Emotions, and Dynamism (see Table 5). Despite being developed from a model for work performance rather than personality, the OPQ provides comprehensive coverage of the personality domain. Many research studies have confirmed a general correspondence of the OPQ32n factor structure with the Five Factor Model (FFM) of personality, with at least five OPQ scales indicating each of the 'Big Five' (Bartram et al., 2006). In addition to the core 32 scales, the OPQ32n captures the tendency to report 'unlikely virtues' via a Social Desirability scale. Extensive information on the psychometric properties of the OPQ32n including reliability, construct and criterion-related validity is available in the OPQ technical manual (Bartram et al., 2006).

The OPQ32n consists of 230 statements of behavior or preference, to which candidates respond using 5 ordered categories: strongly disagree ~ disagree ~ unsure ~ agree ~ strongly agree. Each scale is measured by either 6 or 8 items; item polarity within scales is fully balanced, with half of the items being positive indicators and half negative. Only scale sum scores (without item-level information) were available for this analysis.

For ease of interpretation, all OPQ scale scores were rescaled from sum scores into centered average item scores by dividing each scale score by the respective number of items (6 or 8) and centering on the rating scale midpoint. The resulting scores are comparable across scales and provide a direct reference to the 5-point rating scale, ranging from –2 (as if a respondent 'strongly disagreed' with all positive items and 'strongly agreed' with all negative items) to 2 (as if a respondent 'strongly agreed' with all positive items and 'strongly disagreed' with all negative items), with the theoretical midpoint of 0 (as if a respondent was 'unsure' about all items).

Aptitude. Two tests from the Critical Reasoning Test Battery published by SHL were used to measure candidates' verbal and numerical aptitude. According to the publisher descriptions, Verbal Evaluation (VC1.1) measures 'the ability to understand and evaluate the logic of arguments presented in written passages'; Interpreting Data (NC2.1) measures 'the ability to make correct decisions or inferences from numerical data'. Psychometric properties of these tests are also well documented (SHL, 1991). Only number-correct scores were available to us; we standardized them for ease of interpretation.

Analysis

Fitting alternative measurement models. We adopted the five-factor model (FFM) as a framework for all analyses, as it has a long tradition of use with the OPQ family. Table 5 shows a map of OPQ scales to the Big Five based on EFA of the original UK and US standardization samples described in the OPQ32 Technical Manual (2006). An OPQ scale was included as an indicator of a Big Five factor if its standardized loading was over 0.32 in magnitude in both samples. This map is consistent with the core set of 25 scales used for computing composite Big Five scores but more inclusive, with all 32 OPQ scales mapped and some indicating multiple factors; for instance, Affiliative indicates both Extraversion and Agreeableness.

To assess the appropriateness of a five-factor solution to our data, and also to provide a benchmark for comparison with all confirmatory models based on the Big Five mapping, we first conducted an exploratory factor analysis (EFA), fitting five- and six-factor exploratory models. We used the ML estimator with robust standard errors in Mplus 8.2 (Muthén & Muthén, 2017) to fit the exploratory and the following confirmatory models: (1) **CFA** model measuring Extraversion, Agreeableness, Conscientiousness, Emotional Stability and Openness by OPQ scales as mapped; (2) **IEF** model with the five-factor structure as before plus an ‘ideal-employee’ factor influencing all the OPQ scales and orthogonal to the Big Five factors; (3) **FMA** model with 2 latent classes, each underlain by its own measurement model – the five-factor structure for ‘real’ profiles, and a single factor for ‘ideal’ profiles, as described by Equation (4); and, finally, (4) **F-GoM** model with 2 classes, allowing grade of membership in ‘real’ and ‘ideal’ profiles as regulated by the overall propensity to edit, and with the factor structures for ‘real’ and ‘ideal’ profiles as in the FMA model. The F-GoM model required 7 dimensions of integration, and was estimated using 15,000 random quadrature nodes at which the integrand was evaluated. We note that the OPQ Social Desirability scale was not included in any of the models, but used later for validation.

Testing whether scale importance predicts the propensity to edit. The basic F-GoM model fitted to these data assumed that OPQ scales get edited equally often, which is expressed by equal intercepts α_i in Equation (6), $\alpha_i = \alpha$. To test whether the probability of editing of particular scales actually depends on their importance to selection criteria, we used information from the Occupational Information Network website www.onetonline.org (National Center for O*NET Development, n.d.). Specifically, we used published data on the importance of the so-called Work Styles – 16 personal attributes mapped to success in specified job roles. We extracted the work styles data for O*NET occupations specified in our dataset, including Accountants and Auditors (SOC Code 13-2011), Logisticians (SOC Code 13-1081), Financial Analysts (SOC Code 13-2051), Tax Preparers (SOC Code 13-2082), and Marketing Analysts (SOC Code 13-1161). We summarized the importance of the 16 work styles for these occupations under 3 ordered categories (0=less relevant; 1=important; 2=critical). The work styles **critical** (coded 2) to success in analytical roles were: Analytical thinking, Attention to detail, Achievement/Effort and Dependability. The work styles **important** (coded 1) for success in the target roles were: Independence, Initiative, Integrity, Persistence, Innovation, Adaptability/Flexibility, and Stress Tolerance. The remaining work styles were **less relevant** (coded 0). Finally, we mapped the Work Styles to the OPQ32 scales, resulting in the following importance grades. OPQ scales Data Rational, Evaluative, Detail Conscious, Conscientious, Rule Following and Achieving were graded ‘critical’ (2). OPQ scales Controlling, Independent Minded, Socially Confident, Democratic, Behavioral, Conventional, Innovative, Variety Seeking, Forward Thinking, Relaxed, Worrying, Tough Minded and Vigorous were graded ‘important’ (1). The rest of OPQ scales were graded ‘less relevant’ (0).

With these objective importance criteria, we tested whether more *important* attributes are associated with more editing of the respective scales; that is, whether more responses on that scale will fall into the ‘ideal’ class (C*ij* = 1). The *direction* of scale desirability is irrelevant at the Edit stage (of course, it is relevant at the Select stage – whether higher or lower scores are selected). For example, OPQ Relaxed could be desirable for success and OPQ Worrying undesirable for success; however, both scales are important to consider when selecting an applicant – therefore, we expect more editing in regards to both attributes.

Formally, we examined whether scale importance x_i is positively associated with the propensity to edit scale *i*; or, equivalently, with the ‘ideal’ class membership. To test this hypothesis, we entered the importance grades in F-GoM model as within-level covariates x_i of the latent class variable *c*, as illustrated in Figure 4. According to our hypothesis, the coefficient β in Equation (7) is expected to be positive and significant.

Validation of between-person effects. Although our main goal was to validate the proposed F-GoM model, we validate and briefly discuss the CFA, IEF and FMA models too, as they are popular choices in applications (e.g. A. Brown et al., 2017; Klehe et al., 2012; Pavlov et al., 2019; Zickar et al., 2004). In what follows, we detail the validation strategy for the F-GoM model, and at the end briefly explain how these strategies were applied to the alternative models.

To validate the F-GoM capability in identifying **between-person effects**, including the substantive Big Five factors $\theta^{(R1)}$ to $\theta^{(R5)}$, the tendency to edit responses $\theta^{(E)}$ and the tendency to select extreme responses $\theta^{(S)}$, we used the following observed person-level covariates: gender (men vs. women); Verbal VC11 and Numerical NC21 tests, and OPQ Social Desirability scale – a traditional marker for faking. The validation logic was that if a model is successful in assigning variance to relevant sources – substantive or presentational factors and/or classes, then the between-person effects should exhibit the following theoretically expected relationships with external variables. For the substantive Big Five factors, relationships with covariates should be similar to those observed in low-stakes research samples. For tendencies to edit or to select extreme responses, relationships with covariates should support our assumptions about the nature of these tendencies. Finally, for the assignment of particular OPQ scale scores to ‘real’ or ‘ideal’ classes, convergent correlations with similar constructs (objectively measured) should be high for retrieved responses and low for selected responses. Specific expectations are described below and also summarized in Table 6.

For **gender**, based on the results for OPQ32n US Standardization sample summarized in terms of the Big Five, which are reported in the OPQ Technical Manual, we expected to see men score higher than women on Emotional Stability and Openness; and women score higher than men on Agreeableness (Bartram et al., 2006, p. 195).

For **verbal ability**, based on its conceptual concordance with the tendency to critically evaluate information captured by OPQ Evaluative, we expected the VC11 score to have a positive direct effect on the *retrieved* OPQ Evaluative score but not on the *selected* score, controlling for all the latent tendencies.

For **numerical ability**, based on its conceptual concordance with preference for quantitative analysis (captured by OPQ Data Rational), we expected the NC21 score to have a direct positive effect on the *retrieved* OPQ Data Rational score but not on the *selected* score, controlling for all the latent tendencies. Translating the above hypotheses into the language of model parameters, we expected to see positive effects of objectively assessed verbal and numerical abilities on the respective self-reported attributes in the ‘real’ class but not in the ‘ideal’ class.

Finally, based on the extensive research of **social desirability** scales (McCrae & Costa, 1983; Ones et al., 1996), we expected the OPQ Social Desirability score to relate to both substance and style. For substantive Big Five factors, based on the meta-analysis presented by Ones et al. (1996), we expected the OPQ SDes score to relate positively to Emotional Stability, Conscientiousness and Agreeableness, but not to Extraversion or Openness. For factors capturing faking behavior, we expected the OPQ SDes score to relate positively to the tendencies to Select extreme profiles and Edit responses in F-GoM model, to the general factor in IEF model, and to the 'ideal' class common factor in FMA model.

To test the hypotheses pertaining to the between-person level, all validation covariates were estimated simultaneously using the F-GoM model. The latent variables $\theta^{(R1)}$ to $\theta^{(R5)}$, $\theta^{(S)}$ and $\theta^{(E)}$ were regressed on the person level covariates as illustrated in Figure 4. For verbal and numerical tests, direct paths to OPQ Evaluative and OPQ Data Rational were also modeled, and these effects were allowed to vary in the 'real' and 'ideal' classes. We repeated these analyses for the CFA, IEF and FMA models.

INSERT FIGURE 4 ABOUT HERE

Results

Descriptive statistics. The OPQ32 scale scores were more strongly inter-correlated than reported for research samples, with 23.2% of variance explained by the first component in the current sample compared to 17.6% in the US Standardization sample (Bartram et al., 2006). Panel A in Figure 5 presents box plots (quartiles and outliers) of the observed OPQ scores. It can be seen that scales universally desirable in the workplace, such as Conscientious, and scales specifically relevant to success in analytical jobs, such as Evaluative, yielded high medians and negatively skewed distributions. Scales of ambiguous desirability, such as Competitive, yielded distributions that were more symmetrical. Overall, the evidence so far suggests the presence of impression management, with targeting specific scales rather than all scales. In subsequent analyses, we investigate the extent to which these distributions are mixtures of 'real' levels of applicants' personal styles and levels seen as 'ideal' to put on a job application.

INSERT FIGURES 5 AND 6 ABOUT HERE

Goodness of fit of alternative measurement models. An exploratory factor analysis yielded the scree plot presented in Figure 6, generally supporting a five-factor structure. Table 3 summarizes goodness of fit for the alternative models to the 32 OPQ scale scores. The **five-factor EFA** model fitted adequately according to RMSEA and SRMR, both under .08; however, CFI missed the suggested threshold of .90 for adequate fit (Hu & Bentler, 1999). A target rotation to the hypothesized Big Five structure yielded a solution with significant and strong loadings where expected, but also many smaller non-hypothesized cross-loadings. The **six-factor EFA** model fitted substantially better than the five-factor EFA model according to AIC and BIC, and its CFI was just over .90. A target rotation of the six-factor model to a bifactor structure yielded a solution with most OPQ scales loading on the general factor, and fewer cross-loadings in the Big Five part. In the following, this factor model serves as a benchmark for the other models with the restricted set of Big Five indicators.

The most restricted model, the five-factor **CFA model** fitted the data adequately according to RMSEA but was not acceptable according to CFI and SRMR. Modification indices suggested that many non-hypothesized cross-loadings are needed to improve fit. The model yielded significant positive correlations between all five factors, ranging from .23 to .69 (see Table 4, above diagonal).

Adding a general factor to the Big Five in the **IEF model** improved the fit substantially according to AIC and BIC; and with both RMSEA and SRMR well under .08, indicating adequate fit (see Table 3). The latent factor correlations now ranged from moderate negative −.38 to large positive .52 (see Table S1 in the Supplement). The IE factor yielded loadings corresponding to the direction of scale desirability; for instance, unstandardized loading –0.33 for Worrying, representing the expected OPQ scale point change per one standard deviation change on the IE factor.

INSERT TABLES 3 AND 4 ABOUT HERE

Many researchers restricting their search to factor models would probably stop here and adopt the IEF model as it appears to fit the data well enough. However, we continued to explore the data with a factor mixture analysis and found that the **FMA model** with the hypothesized Big Five structure in the 'real' class and just one factor in the 'ideal' class fitted decisively better than the IEF model according to AIC; however, BIC favored the less-parameterized IEF model (see Table 3). More applicant profiles were classified as 'real' than 'ideal' (59% and 41% respectively). The OPQ scale means in the 'real' and 'ideal' classes differed in the expected directions of scale desirability; for example, the Worrying mean was 0.44 scale points lower in the 'ideal' class. The latent Big Five correlations were mostly positive, ranging between –.04 and .56, which overall are slightly higher than observed for the IEF model but lower than for the CFA model (see Table S1 in the Supplement). These substantive factors yielded loadings in the expected directions according to the OPQ scale mapping. The presentational factor $\theta^{(S)}$ in the 'ideal' class yielded loadings in the direction of scale desirability, for instance, –0.25 for OPQ Worrying (unstandardized).

Finally, the **F-GoM model** in its most basic and constrained form with all scales assumed equally 'easy' to edit (i.e., all intercepts in Equation (6) set equal, $\alpha_i = \alpha$), required the estimation of only 12 more parameters than the FMA model. The improvement in fit, however, was greater than the improvement from the baseline CFA to the FMA model (which doubled the number of parameters estimated; see Table 3). Moreover, this model fitted far better than the most relaxed six-factor EFA model with more parameters.

F-GoM model results and interpretation. The basic F-GOM model yielded easily interpretable results. The overall prevalence of 'real' scale reports was actually slightly lower than 'ideal', 49% versus 51% (note that class prevalence here is calculated as the proportion of all person*scale reports). Panel B in Figure 5 shows box plots of the observed OPQ scores by most likely class – 'real' or 'ideal'. The plot shows that the scale means differed between the classes in the expected directions of desirability for analytical jobs. These differences were larger in magnitude than in the FMA model, up to around 1 rating scale point. For example, the mean for Worrying was 1.19 scale points lower in the 'ideal' class. Figure 7 shows the distributions of observed scores by likely class membership for selected OPQ scales. It illustrates that some scales yielded excellent separation of classes and large mean differences, for instance Worrying (panel A). Other scales showed different means but substantial overlaps in the class distributions; for instance, Evaluative (panel B). Competitive (panel C) showed similar means in the 'real' and 'ideal' classes but while the 'real' scores spanned a large range, the 'ideal' scores were distributed compactly around their mean.

INSERT TABLE 5 ABOUT HERE

Table 5 contains the unstandardized factor loadings of the OPQ scales for the substantive Big Five factors and the tendency to select extreme scores $\theta^{(S)}$ of the F-GoM model. The pattern of loadings corresponds to the conceptual meaning of these factors. For example, OPQ Innovative, which is a desirable attribute indicating Openness, loaded strongly on both $\theta^{(S)}$ and $\theta^{(R\text{-}Open)}$. For another example, Rule Following, a desirable attribute in organizational settings, which positively indicate Conscientiousness and negatively Openness, loaded positively on $\theta^{(S)}$ and $\theta^{(R\text{-}Cons)}$, and negatively on $\theta^{(R\text{-}Open)}$. Finally, Competitive, a negative indicator of Agreeableness with ambivalent Desirability, loaded negatively on $\theta^{(R\text{-}Agre)}$ and at zero on $\theta^{(S)}$.

Moving to the relationships between the latent factors, the Big Five correlations ranged between –.21 and .74 (see Table 4). We note that this range is comparable to the one from the IEF and FMA models. All the Big Five factors except Agreeableness had small to medium positive correlations with the tendency to edit $\theta^{(E)}$ (see Table 4), with Extraversion yielding the strongest association (r = .55), which is to be expected considering the inclusion of assertive and self-promoting attributes (OPQ Persuasive and reversed OPQ Modest). Moreover, the Big Five factors correlated positively with $\theta^{(S)}$, indicating that applicants with higher level of attributes tend to produce highly desirable profiles when they edit. Together with Extraversion, Conscientiousness yielded the strongest associations with this tendency (both r = .50), which can be explained by a strong element of achievement orientation (OPQ Achieving).

The propensity to edit $\theta^{(E)}$ is an important person-level variable to interpret. The estimated variance of $\theta^{(E)}$ was highly significant but small at 0.50, confirming that partial (intermittent) rather than exclusive class memberships were indeed typical in this sample (Asparouhov & Muthen, 2007). Finally, the small correlation –0.11 between $\theta^{(E)}$ and the presentational factor $\theta^{(S)}$ was not significant (p = .18), suggesting that the two faking tendencies were largely independent from one another.

Does scale importance predict the probability of scale editing? We tested the role of scale importance using the importance ratings of OPQ scales derived from the O*NET as within-level covariate of decision to edit. This model estimated one more parameter than the basic F-GoM model – the coefficient β as per Equation (7). The model fit improved substantially according to AIC (Δ = 32) and BIC (Δ = 24) (see the penultimate and last columns in Table 3). The coefficient β was positive and significant at 0.44 (*p* < .001), thus supporting our hypothesis that attributes important to job success elicit more editing. The effect equates to an odds ratio of 1.55, which means that for a one-point change in the importance rating (e.g. from 'less relevant' to 'important', or from 'important' to 'critical'), the odds of an applicant editing the attribute increase 1.55 times.

Results of validation of between-person effects. Table 6 summarizes the validation results of the between-person effects using external covariates. This table presents the standardized regression coefficients of the latent factors and observed scores for selected OPQ scales on the between-person covariates entered simultaneously in the best-fitting F-GoM model with within-person covariates x_i. Results pertaining to the alternative models – CFA, IEF and FMA are provided in Table S2 of the Supplement.

INSERT TABLE 6 ABOUT HERE

For **gender**, women scored significantly higher on Agreeableness and lower on Emotional Stability and Openness (marginally significant, *p* = .06) than men in the F-GoM model, consistent with our expectations. This was also the case in the other models, with the exception of the IEF model where the effects for Emotional Stability and Openness were insignificant, but the effects of Extraversion and Conscientiousness both positive and significant. Gender did not have any significant effect on tendencies to Edit or to Select extreme responses in the F-GoM model, nor on the common factor in the 'ideal' class in the FMA model. In the IEF model, however, women had significantly lower scores on the ideal-employee factor.

Verbal ability had a positive small effect on OPQ Evaluative in the models without latent classes (CFA and IEF). The mixture models (FMA and F-GoM) differentiated the effect between classes, so that it was higher in the ‘real’ class and lower in the ‘ideal’ class. This supports the validity of the identified classes, and highlights the value of considering latent sub-populations in models of faking. The F-GoM model in particular showed substantial differences in the OPQ Evaluative score in the ‘real’ class, where its standardized path on the Verbal test was .35, and .07 in the ‘ideal’ class, further supporting the validity of the model specification.

Numerical ability had a positive small effect on OPQ Data Rational in the models without latent classes (CFA and IEF). Again, the effect was differentiated between classes in the mixture models (FMA and F-GoM). The F-GoM model showed better differentiation than the FMA model, with the substantial standardized effect .40 in the ‘real’ class, and negligible .09 in the ‘ideal’ class as expected. Neither Verbal nor Numerical test scores had any effect of note on the Big Five, tendency to edit $\theta^{(E)}$ or to produce extreme profiles $\theta^{(S)}$ in the F-GoM model, or any non-substantive factors in the IEF or FMA models.

The traditional marker of faking, the OPQ **Social Desirability** (SDes) scale had substantial (.40 to .45) positive effects on Emotional Stability, Conscientiousness and Agreeableness, and no effects of any note on Extraversion or Openness in the F-GoM model, exactly as expected. The results were similar in the alternative models. The F-GoM model yielded medium-size positive effects of SDes on the tendency to produce extremely desirable profile $\theta^{(S)}$ (.27) and the tendency to edit $\theta^{(E)}$ (.41), again as expected. This result supports the notion that social desirability scales capture both substance and style. The IEF model yielded only a small positive effect (.20) of the SDes scale on the ‘ideal-employee’ factor, and the FMA model yielded a medium effect (.40) on the presentational factor in the ‘ideal’ class.

Discussion of Application Results

This analysis illustrated that the proposed modeling approach is suitable for operational assessment data, even with self-reports recorded only at the scale level. Given the high-stakes assessment context, where response distortions manifest themselves in skewed distributions and inflated correlations between scales, the F-GoM model fitted the observed data much better than the other established models for faking. Specifically, the fit was improved substantially by allowing for a scale-by-scale membership in 'real' or 'ideal' classes compared to the fixed class membership in standard FMA modeling. The fit was improved still further by allowing the probability of editing each particular OPQ scale to differ depending on its importance for job success. Thus, attributes that are more important in analytical jobs according to the O*NET database (National Center for O*NET Development, n.d.) appear to get edited more often than less important attributes by the same individuals.

The F-GoM latent variables and classes validated well against objectively measured demographics and cognitive abilities, and also the traditional marker of faking behavior – a social desirability scale. Thus, the Big Five factors yielded gender differences consistent with those found in the OPQ32n US Standardization sample, which was collected in low-stakes research settings. Moreover, only substantive Agreeableness, Conscientiousness and Emotional Stability were related to the OPQ Social Desirability scale, again consistent with the low-stakes personality research (Ones et al., 1996). These findings support the validity of the Retrieve-stage model. The propensity to Edit and the tendency to Select extreme scores were both strongly related to Social Desirability but not to gender or cognitive abilities, supporting the validity of the Edit and Select stage models. The most encouraging result was that the F-GoM identification of intermittent faking (i.e. scale-specific editing) validated against objective measures since performance on the numerical test was found to be associated with self-reported preference for numerical analysis, but only for reports classified as ‘real’. Misreported interest for numerical analysis, as in the ‘ideal’ class, no longer correlated with the actual ability. We conclude that the F-GoM analysis improved the validity of personality assessments by controlling more effectively for faking effects than the other considered models.

Conclusions and Discussion

This paper introduced the Faking-as-Grade-of-Membership (F-GoM) model to allow for intermittent faking in high-stakes assessments, thus overcoming the restrictive assumption that faking behavior is consistent throughout assessment. The basic idea is that each observed response is either ‘real’ (representing the respondent’s retrieved, or perceived own standing on each attribute in question), or ‘ideal’ (representing the strategic response that portrays the desired picture of the respondent in the eyes of the assessor). Thus, respondents may decide to report honestly on the attributes that they are good at, or attributes they do not deem important enough as selection criteria; on the other hand, they may feel compelled to present an ‘ideal’ image on important attributes they feel they are lacking. By providing a mechanism for switching between ‘real’ and ‘ideal’ responses, the proposed method accommodates the process of self-censorship by which people consider and potentially edit their retrieved responses before reporting. Similar processes have been suggested in the context of responding to sensitive survey questions (e.g. Böckenholt, 2014; Tourangeau & Yan, 2007), socially desirable responding (Holtgraves, 2004) and lying (Walczyk et al., 2005). By providing separate structural models for ‘real’ and ‘ideal’ responses, the proposed method also accounts for the heterogeneity of respondent behavior, strongly indicated in previous research on test taker goals (Kuncel et al., 2011) and activation mechanisms (Robie et al., 2007), and identified in actual response data (Zickar et al., 2004).

Intermittent faking enabled by the proposed approach is not only plausible theoretically; it is well supported as an important mechanism in empirical data – by vastly improved model fits as well as strong validity gains in operational high-stakes assessments compared to current models which assume that person faking behavior is consistent throughout the assessment. Specifically, the fixed-class factor mixture models (Zickar et al., 2004), which can be considered a special case of the proposed F-GoM model when the tendency to Edit has infinite variance, yielded demonstrably worse model-data fit in the reported application and weaker validity of model-predicted scale scores. Thus, the correlation between self-reported strength in data analysis and objectively measured numerical ability was weaker in the member-exclusive 'real' class of the FMA model than in the 'real' class of the F-GoM model with partial membership. Therefore, we believe that the partial membership in 'real' and 'ideal' response profiles is the main strength of the proposed approach. Other strengths include:

1) The approach allows for separate measurement models for 'real' and 'ideal' responses that can be specified in multiple ways depending on the context;

2) The model is implemented as a multilevel mixture, allowing the addition of class-specific explanatory covariates at the item level and the person level;

3) The model can be used with either item-level or scale-level data, with the latter allowing for flexible data analyses when item-level data are either unavailable or unwieldy (too many scales are measured);

4) The model can be used readily in applied work by utilizing general-purpose SEM software.

With regard to (1), the measurement models for 'real' and 'ideal' responses do not have to be, and generally should not be, the same. In the present paper we use linear factor analysis models (with logit or probit links for modeling categorical responses as appropriate), with as many factors as measured scales in the ‘real’ class and only one factor controlling for the tendency to produce extremely desirable profiles in the ‘ideal’ class. These may be reasonable for many applications, but the choice will be determined by the measurement instrument analyzed and the assessment context. For instance, ideal-point models may also be considered to describe the retrieved or selected responses.

With regard to (2), one example is the use of external covariates of the item- or scale-specific decisions to edit. As was suggested earlier, decisions to edit particular responses may be influenced not only by individual differences in the propensity to edit, but also by the properties of items to which people respond. These properties may be either known (e.g., expert rated) or estimated from the data. For instance, items that are central to the selection criteria may be edited more often than items that have less relevance to the selection criteria. The Application presented in this paper utilized objective importance ratings derived from O*NET as covariates of scale-specific editing. In future studies, it may be explored whether candidate assessments of what is important for the job are also useful in predicting decisions to edit.

With regard to (3), modelling at the scale rather than at the item level simplifies estimation and allows for a richer set of analyses that may be otherwise impossible for computational reasons. Scale scores can be computed by whichever method or protocol normally used for scoring, and can be entered in the analysis to model faking behavior in practical settings. In such analyses, the goal is to determine to what extent each particular scale score is likely to be from the test taker’s ‘real’ or ‘ideal’ personality profiles. The Application demonstrated that analyses of scale scores of multi-scale questionnaires can provide reasonable control for intermittent faking, and identify retrieved (or ‘real’) and selected (or ‘ideal’) scores and their governing factors when the number of scales is large, with an a-priori factor structure at the Retrieve stage. It may be more difficult to analyze scale scores without a clear factor structure, or with some scales not fitting the structure. Although it is possible to have an unrestricted covariance structure at the Retrieve stage, this will add as many dimensions of integration as there are units of analysis (here, scales). A Bayesian estimation approach could provide a solution to this problem.

Although the Application demonstrated that analyzing scale scores can be very useful when the conditions are right, ignoring the item level decisions may lead to distortions yet unknown. In this paper we argued that scale scores analyzed with F-GoM can be thought of as proxies for self-reported standings on attributes the test is designed to measure. These could be Retrieved or Selected without assuming any particular mechanism for aggregating the decisions taken at the item level. This approach is most appropriate for narrow personality constructs, usually measured by items with similar content, because in this case self-reports on the constituent items are likely to aggregate to a score that captures a self-rating on the construct. However, more research is needed to identify the exact conditions under which the postulated decision stages at the scale level correspond to the decision stages at the item level.

Finally, with regard to (4), the practical implementation of the proposed approach as a multilevel mixture model makes it easier to capitalize on its strength in applied work. Going forward, the F-GoM approach should be explored for obtaining estimates of the relationship between work performance and (faking-controlled) personality measures. Previously, much research in this area has confounded the substantive and presentational factors or has attempted to adjust for faking using techniques with known problems, such as manifest indices (social desirability or lie scales). Moreover, not every latent variable technique is effective in separating the substantive and presentational effects; for instance, the general-factor approach to modeling over-reporting can over-extract substantive variance thus distorting research results. This was evident from the Simulation study, which uncovered the tendency of the IEF model to under-estimate true variance shared by responses (illustrated in Figure 2); and to some extent from the Application, which demonstrated poorer performance of the IEF model in the validation against external variables.

We hope that the suggested approach will prove useful in informing the design of future assessments aimed at minimizing the effects of faking behavior and at maximizing validity – both construct and criterion-related. It may also help in designing and further developing assessment measures that are fake proof. One possibility is to embed F-GoM analyses in experimental studies at questionnaire trialing stages to establish what types of items/scales are most affected, and whether any interventions are effective in preventing or reducing the effects of faking. Further progress will depend on improvements in our understanding of the response process and its underlying tendencies which ultimately will make measurements of faking behavior obsolete. We are looking forward to advances in this area.

References

Ariely, D. (2012). *The (honest) truth about dishonesty*. Harper.

Asparouhov, T., & Muthen, B. (2007). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27–51). Information Age Publishing, Inc.

Asparouhov, T., & Muthén, B. (2017). *Using Mplus individual residual plots for diagnostics and model evaluation in SEM* (No. 20; Mplus Web Notes, Issue 20).

Bartram, D., Brown, A., Fleck, S., Inceoglu, I., & Ward, K. (2006). *OPQ32 Technical Manual*. SHL Group.

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

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

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

Brown, E. A., & Grice, J. W. (2011). One is enough: Single-item measurement via the dynamic analog scale. *SAGE Open*, *1*(3), 1–10. https://doi.org/10.1177/2158244011428647

Clark, S. L., Muthén, B., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: Two examples concerning the structure underlying psychological disorders. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4). https://doi.org/10.1080/10705511.2013.824786

Haaland, D., & Christiansen, N. D. (1998). Departures from linearity in the relationship between applicant personality test scores and performance as evidence of response distortion. *22nd Annual International Personnel Management Association Assessment Council Conference*. https://pdfs.semanticscholar.org/70ce/2add187e210e32883cc8a52c74d1a003bd82.pdf

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, L., & Bentler, P. M. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

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

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. 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. *Personnel Psychology*, *62*(2), 201–228. https://doi.org/10.1111/j.1744-6570.2009.01136.x

Leite, W. L., & Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research*, *45*(January 2015), 271–293. https://doi.org/10.1080/00273171003680245

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

McDonald, R. P. (1999). *Test Theory: A unified treatment* (2011th ed.). Erlbaum.

Muthen, B. O. (2007). Latent variable hybrids: Overview of old and new models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 1–24). Information Age Publishing, Inc.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus User's Guide. Eighth Edition.* Muthén & Muthén.

National Center for O*NET Development. (n.d.). *O*NET Online*. Retrieved November 29, 2019, from https://www.onetonline.org

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

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

Pavlov, G., Maydeu-Olivares, A., & Fairchild, A. J. (2019). The effects of applicant faking on forced-choice and Likert scores. *Organisational Research Methods*, *22*(3), 710–739. https://doi.org/https://doi.org/10.1177/1094428117753683

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

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

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

SHL. (1991). *Critical Reasoning Test Battery: Manual and User's Guide*. SHL Group.

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

Walczyk, J. J., Schwartz, J. P., Clifton, R., Adams, B., Wei, M., & Zha, P. (2005). Lying person-to-person about life events: A cognitive framework for lie detection. *Personnel Psychology*, *58*(1), 141–170. https://doi.org/10.1111/j.1744-6570.2005.00484.x

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

Table 1. *Simulation Results of F-GoM Model: Recovery of Measurement (Item) Parameters*

Class 'real' (Retrieve stage)								*Class 'ideal'(Select stage)*							
		N = 1000			N=500					N = 1000			N=500		
	True	Mean Est.	Mean SE	95% cover	Mean Est.	Mean SE	95% cover		True	Mean Est.	Mean SE	95% cover	Mean Est.	Mean SE	95% cover
Intercepts															
$\mu_{1,0}$	**0**	0.003	0.097	0.96	-0.002	0.141	0.95	$\mu_{1,1}$	**2**	2.003	0.066	0.96	1.995	0.097	0.97
$\mu_{2,0}$	**0**	-0.001	0.105	0.96	-0.005	0.153	0.94	$\mu_{2,1}$	**1**	1.000	0.081	0.95	0.994	0.118	0.93
$\mu_{3,0}$	**0**	0.001	0.098	0.96	-0.005	0.145	0.96	$\mu_{3,1}$	**1**	1.004	0.069	0.96	1.009	0.100	0.95
$\mu_{4,0}$	**0**	0.001	0.094	0.95	-0.003	0.140	0.96	$\mu_{4,1}$	**0**	-0.002	0.070	0.96	0.000	0.102	0.94
$\mu_{5,0}$	**0**	0.008	0.091	0.96	0.005	0.133	0.96	$\mu_{5,1}$	**-1**	-1.003	0.053	0.95	-1.003	0.076	0.96
$\mu_{6,0}$	**0**	-0.001	0.094	0.95	0.005	0.138	0.96	$\mu_{6,1}$	**2**	1.999	0.066	0.96	1.995	0.097	0.94
$\mu_{7,0}$	**0**	-0.002	0.092	0.96	-0.004	0.135	0.96	$\mu_{7,1}$	**1**	0.999	0.073	0.96	1.002	0.107	0.94
$\mu_{8,0}$	**0**	0.004	0.089	0.96	0.004	0.132	0.97	$\mu_{8,1}$	**1**	1.001	0.064	0.94	1.004	0.094	0.94
$\mu_{9,0}$	**0**	0.000	0.087	0.97	0.006	0.130	0.96	$\mu_{9,1}$	**0**	-0.005	0.064	0.95	-0.005	0.093	0.97
$\mu_{10,0}$	**0**	0.001	0.087	0.96	0.007	0.129	0.96	$\mu_{10,1}$	**-1**	-1.004	0.052	0.96	-1.002	0.077	0.96
Factor loadings:		$\theta^{(R1)}$ *and* $\theta^{(R2)}$						*Factor loadings:*		$\theta^{(S)}$					
λ_1^{R1}	**1**	1.002	0.087	0.96	1.004	0.129	0.96	s_1	**0.7**	0.700	0.064	0.95	0.701	0.095	0.96
λ_2^{R1}	**1**	0.996	0.093	0.95	0.999	0.139	0.96	s_2	**0.7**	0.697	0.071	0.94	0.698	0.104	0.95
λ_3^{R1}	**1**	0.994	0.091	0.97	0.993	0.136	0.96	s_3	**0.5**	0.500	0.065	0.97	0.496	0.095	0.94
λ_4^{R1}	**1**	0.999	0.092	0.95	0.996	0.137	0.96	s_4	**0.5**	0.502	0.072	0.94	0.500	0.105	0.95
λ_5^{R1}	**1**	1.001	0.087	0.96	1.003	0.130	0.96	s_5	**0**	0.004	0.059	0.96	0.008	0.087	0.96
λ_6^{R2}	**1**	0.999	0.090	0.97	0.999	0.136	0.97	s_6	**0.7**	0.693	0.065	0.96	0.697	0.096	0.94
λ_7^{R2}	**1**	0.993	0.093	0.95	0.989	0.139	0.96	s_7	**0.7**	0.698	0.070	0.95	0.698	0.102	0.95
λ_8^{R2}	**1**	0.998	0.092	0.95	0.987	0.138	0.96	s_8	**0.5**	0.493	0.065	0.96	0.488	0.095	0.96
λ_9^{R2}	**1**	0.996	0.094	0.97	0.998	0.141	0.97	s_9	**0.5**	0.500	0.067	0.96	0.496	0.099	0.96
λ_{10}^{R2}	**1**	0.996	0.092	0.95	0.994	0.137	0.96	s_{10}	**0**	0.000	0.059	0.96	-0.005	0.086	0.96
Residual variances															
var($\varepsilon_{1,0}$)	**1**	0.988	0.143	0.96	0.972	0.210	0.94	var($\varepsilon_{1,1}$)	**0.5**	0.492	0.074	0.94	0.494	0.108	0.94
var($\varepsilon_{2,0}$)	**1**	0.993	0.147	0.94	0.980	0.215	0.93	var($\varepsilon_{2,1}$)	**0.5**	0.489	0.080	0.94	0.482	0.114	0.93
var($\varepsilon_{3,0}$)	**1**	0.997	0.142	0.95	0.979	0.207	0.93	var($\varepsilon_{3,1}$)	**0.5**	0.492	0.066	0.95	0.485	0.095	0.93
var($\varepsilon_{4,0}$)	**1**	0.982	0.131	0.94	0.957	0.196	0.92	var($\varepsilon_{4,1}$)	**0.5**	0.497	0.074	0.93	0.495	0.106	0.90
var($\varepsilon_{5,0}$)	**1**	0.981	0.129	0.95	0.971	0.191	0.95	var($\varepsilon_{5,1}$)	**0.5**	0.497	0.053	0.95	0.487	0.076	0.95
var($\varepsilon_{6,0}$)	**1**	0.980	0.156	0.93	0.975	0.231	0.93	var($\varepsilon_{6,1}$)	**0.5**	0.497	0.073	0.95	0.493	0.106	0.96
var($\varepsilon_{7,0}$)	**1**	0.992	0.150	0.94	0.974	0.220	0.94	var($\varepsilon_{7,1}$)	**0.5**	0.489	0.079	0.94	0.485	0.113	0.94
var($\varepsilon_{8,0}$)	**1**	0.982	0.148	0.96	0.973	0.219	0.96	var($\varepsilon_{8,1}$)	**0.5**	0.501	0.067	0.94	0.502	0.097	0.96
var($\varepsilon_{9,0}$)	**1**	0.978	0.149	0.96	0.966	0.221	0.95	var($\varepsilon_{9,1}$)	**0.5**	0.501	0.072	0.95	0.499	0.104	0.95
var($\varepsilon_{10,0}$)	**1**	0.993	0.150	0.95	0.978	0.222	0.93	var($\varepsilon_{10,1}$)	**0.5**	0.492	0.053	0.93	0.490	0.077	0.94

Table 2.

Simulation Results of F-GoM Model: Recovery of Structural Parameters

		N=1000			N=500		
	True	Mean Est.	Mean SE	95% cover	Mean Est.	Mean SE	95% cover
Factor covariances							
cov($\theta^{(R1)}$,$\theta^{(R2)}$)	**0.3**	0.296	0.060	0.95	0.288	0.091	0.97
cov($\theta^{(R1)}$,$\theta^{(S)}$)	**0.5**	0.508	0.073	0.95	0.511	0.107	0.96
cov($\theta^{(R1)}$,$\theta^{(E)}$)	**0.5**	0.503	0.095	0.97	0.491	0.143	0.96
cov($\theta^{(R2)}$,$\theta^{(S)}$)	**0**	0.001	0.077	0.95	–0.001	0.116	0.96
cov($\theta^{(R2)}$,$\theta^{(E)}$)	**0**	–0.004	0.096	0.96	–0.010	0.146	0.97
cov($\theta^{(S)}$,$\theta^{(E)}$)	**0**	0.009	0.105	0.97	0.007	0.157	0.96
Propensity to edit $\theta^{(E)}$							
α	**0**	–0.004	0.099	0.97	0.002	0.142	0.95
var($\theta^{(E)}$)	**1**	0.984	0.197	0.93	0.984	0.197	0.93

Table 3.

Application: Goodness of Fit for Alternative Measurement Models

Model type	**EFA**	**EFA**	**CFA**	**IEF**	**FMA**	**F-GOM** (α_i=α)	**F-GOM** (α_i=α+βx_i)
# classes	1	1	1	1	2	2	2
# factors	5	6	5	6	5 / 1	5 / 1	5 / 1
#parameters	214	241	112	143	209	221	222
Loglikelihood	-17360	-17217	-18128	-17668	-17566	-16762	-16745
AIC	35148	34917	36480	35622	35550	33966	33934
BIC	36140	36033	36999	36285	36519	35756	35732
Chi-square (df)	1187(346)	937(319)	2492(448)	1711(417)			
CFI	.88	.91	.70	.81			
RMSEA	.056	.050	.077	.064			
RMSEA 90% CI	.053 .060	.047 .054	.074 .080	.061 .067			
SRMR	.035	.029	.091	.058			

Note. Loglikelihood, AIC, BIC and Chi-square values are rounded to the nearest integer. EFA = Exploratory Factor Analysis; CFA = Confirmatory Factor Analysis; IEF = Ideal Employee Factor; FMA = Factor Mixture Analysis; F-GoM = Faking as Grade of Membership.

Table 4.

Application: Estimated Correlations Between Latent Factors in Basic CFA Model and in F-GoM Model With Equal Within-Level Intercepts ($\alpha_i=\alpha$)

	$\theta^{(R\text{-}Extra)}$	$\theta^{(R\text{-}Agre)}$	$\theta^{(R\text{-}Cons)}$	$\theta^{(R\text{-}Emot)}$	$\theta^{(R\text{-}Open)}$	$\theta^{(S)}$
$\theta^{(R\text{-}E)}$ Extraversion		$.43^{***}$	$.59^{***}$	$.65^{***}$	$.69^{***}$	
$\theta^{(R\text{-}A)}$ Agreeableness	.15		$.61^{***}$	$.44^{***}$	$.23^{***}$	
$\theta^{(R\text{-}C)}$ Conscientious.	$.52^{***}$	$.24^{**}$		$.60^{***}$	$.56^{***}$	
$\theta^{(R\text{-}ES)}$ Emot. Stability	$.74^{***}$	.09	$.46^{***}$		$.56^{***}$	
$\theta^{(R\text{-}O)}$ Openness	$.57^{***}$	$-.21^{*}$	$.21^{*}$	$.49^{***}$		
$\theta^{(S)}$	$.50^{***}$	$.33^{***}$	$.50^{***}$	$.46^{***}$	$.46^{***}$	
$\theta^{(E)}$	$.55^{***}$	.01	.18	$.38^{***}$	$.24^{*}$	–.11

Note. Correlations for the CFA model are *above* the diagonal; for the F-GoM model *below* the diagonal. * Values are significant at the .05 level; ** at .01 level; *** at .001 level.

Table 5.

Application: Unstandardized Factor Loadings in F-GOM Model With Equal Within-Level Intercepts ($\alpha_i = \alpha$)

		c = ‘real’					c=’ideal’
Domain	*OPQ32 scale*	$\theta^{(R\text{-}Extra)}$	$\theta^{(R\text{-}Agre)}$	$\theta^{(R\text{-}Cons)}$	$\theta^{(R\text{-}Emot)}$	$\theta^{(R\text{-}Open)}$	$\theta^{(S)}$
Relationships with People	1 Persuasive	0.59					0.34
	2 Controlling	0.53					0.19
	3 Outspoken	0.30	–0.16				0.18
	4 Independent minded		–0.29				0.18
	5 Outgoing	0.50					0.30
	6 Affiliative	0.24	0.28				0.22
	7 Socially Confident	0.33			0.36		0.28
	8 Modest	–0.47					$0.05^{n/s}$
	9 Democratic		0.28				0.22
	10 Caring		0.35				0.27
Thinking Styles	11 Data Rational			0.20			0.21
	12 Evaluative			0.16		0.12	0.30
	13 Behavioral		0.22			0.22	0.31
	14 Conventional					–0.57	$–0.04^{n/s}$
	15 Conceptual					0.40	0.29
	16 Innovative					0.57	0.34
	17 Variety Seeking					0.57	0.18
	18 Adaptable					0.11	–0.16
	19 Forward Thinking			0.45			0.31
	20 Detail Conscious			0.29			0.25
	21 Conscientious			0.38			0.28
	22 Rule Following			0.33		–0.47	0.26
Feelings and Emotions	23 Relaxed				0.38		0.26
	24 Worrying				–0.37		–0.24
	25 Tough Minded				0.41		0.28
	26 Optimistic				0.34		0.28
	27 Trusting		0.27				0.26
	28 Emotionally Controlled	–0.12					–0.16
Dynamism	29 Vigorous			0.24			0.33
	30 Competitive		–0.23				$0.01^{n/s}$
	31 Achieving			0.43			0.29
	32 Decisive		–0.14				0.17

Note. Blanks correspond to loadings fixed to 0 by design. Unstandardized factor loadings correspond to a change on the 5-point OPQ32n response scale per 1 SD change in the latent factor. All values are significant at the .01 level unless marked ‘n/s’.

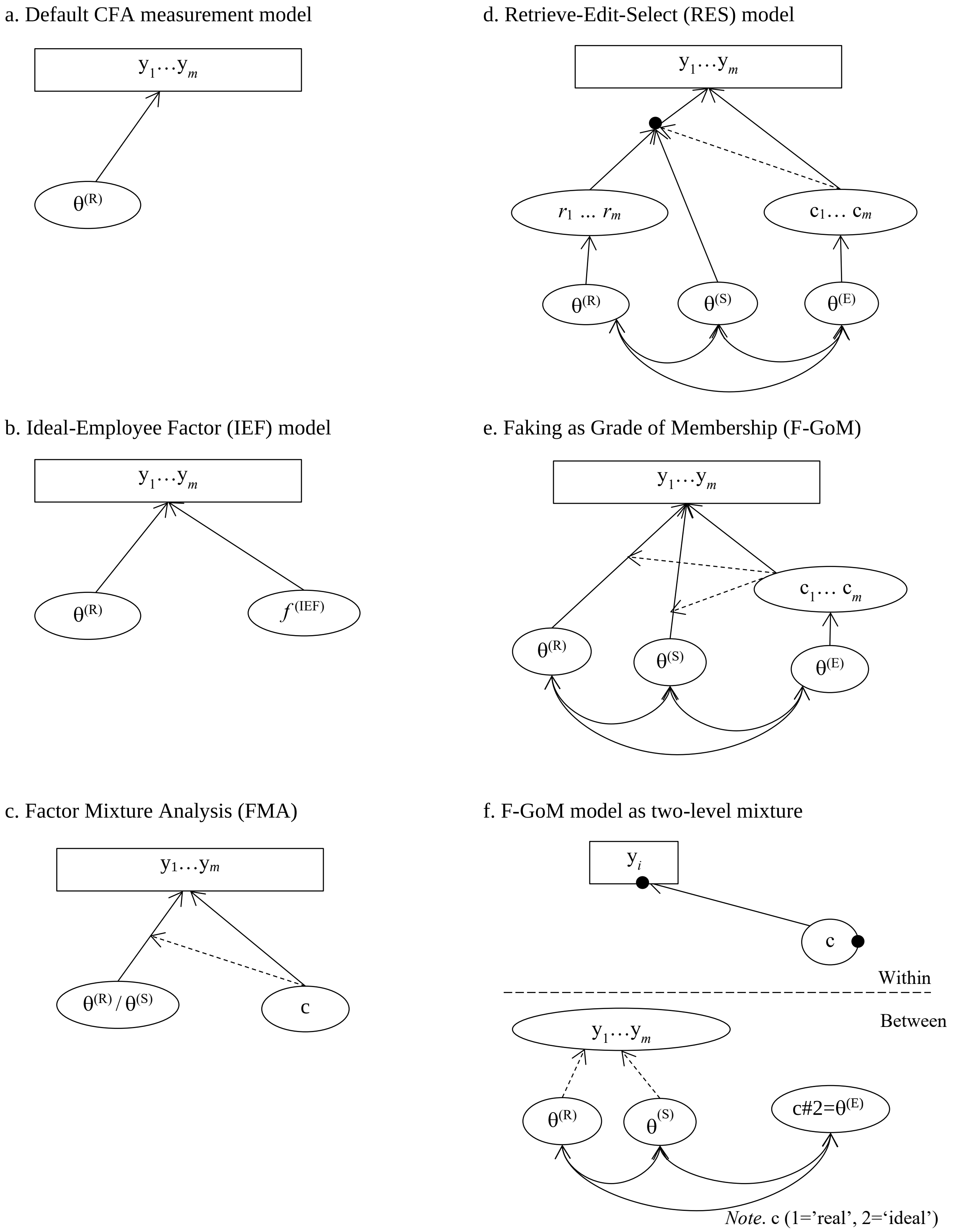
Table 6.

Application: Standardized Between-level Effects on Latent and Observed Variables in F-GOM Model with Within-Level Covariates x_i ($\alpha_i = \alpha + \beta x_i$)

	Covariate			
Dependent variable	Gender (1=women)	Verbal VC11	Numerical NC21	OPQ Social Desirability
Expected effects	Significant for some substantive factors	Positive on Evaluative in 'real' but not 'ideal' class	Positive on Data Rational in 'real' but not 'ideal' class	Significant for both substantive and faking factors
$\theta^{(\text{R-Extraversion})}$	–.05	–.01	–.05	.07
$\theta^{(\text{R-Agreeableness})}$	. 29***	.08	–.01	.40***
$\theta^{(\text{R-Conscientiousness})}$	.04	–.12	.15*	.45***
$\theta^{(\text{R-Emot.Stability})}$	–.18**	.05	.03	.44***
$\theta^{(\text{R-Openness})}$	–.12	.08	.04	.05
$\theta^{(S)}$	–.09	.10	.07	.27***
$\theta^{(E)}$	–.20	–.06	.11	.41***
OPQ Evaluative		.35**/.07		
OPQ Data Rational			.40***/ .09*	

Note. Values before and after forward slash (/) are paths in 'real' and 'ideal' classes, respectively. These coefficients are standardized on the basis of the whole sample (not within classes). * Values are significant at the .05 level; ** at .01 level; *** at .001 level.

Figure 1. *Measurement model diagrams*



a. Default CFA measurement model

b. Ideal-Employee Factor (IEF) model

c. Factor Mixture Analysis (FMA)

d. Retrieve-Edit-Select (RES) model

e. Faking as Grade of Membership (F-GoM)

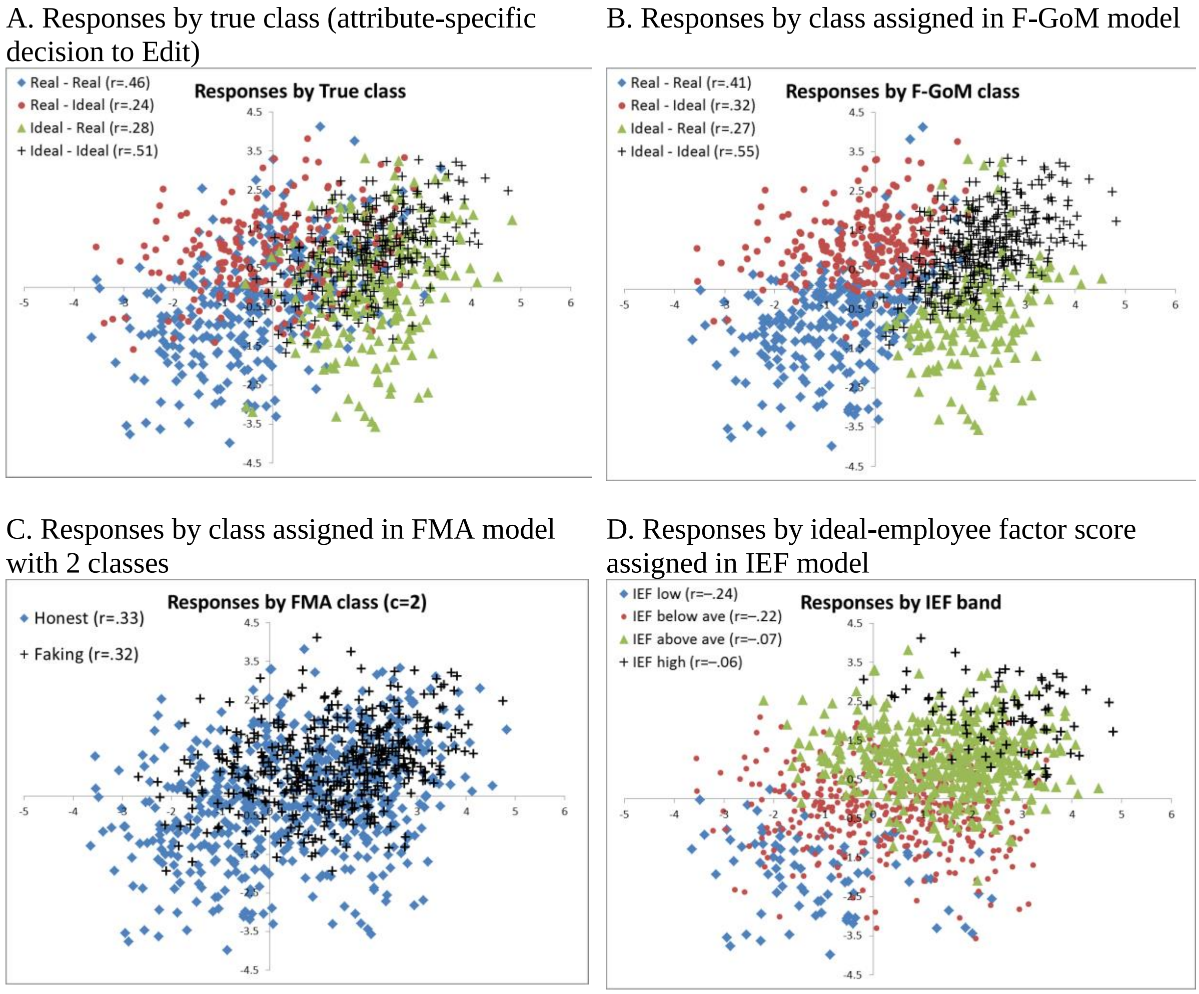
f. F-GoM model as two-level mixture

Note. c (1='real', 2='ideal')

Note. 'c' denotes categorical variables; '*f*', 'θ' and '*r*' denotes continuous variables

Figure 2.

Simulation: Scatterplots of 'observed' responses Y1 and Y2 in Replication 1



A. Responses by true class (attribute-specific decision to Edit)

B. Responses by class assigned in F-GoM model

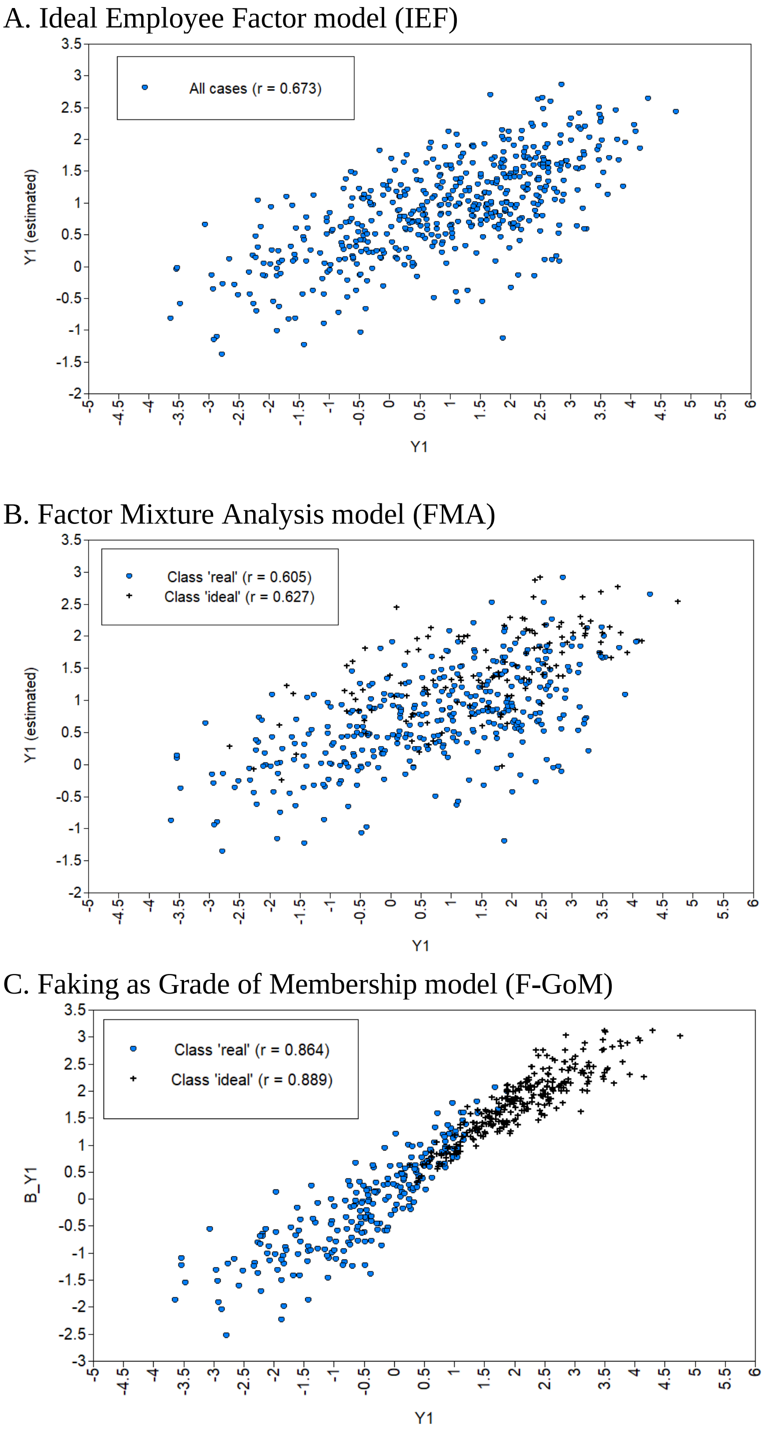
C. Responses by class assigned in FMA model with 2 classes

D. Responses by ideal-employee factor score assigned in IEF model

Note. Markers are set according to: (A) 'true' class; (B) most likely class in F-GoM model; (C) most likely class in FMA model; and (D) most likely score on the IEF factor categorized into four bands. Legends display observed correlations between points within each marker group.

Figure 3.

Simulation: ‘Observed’ Versus Model-Predicted Responses for Y1 in Replication 1


A. Ideal Employee Factor model (IEF)

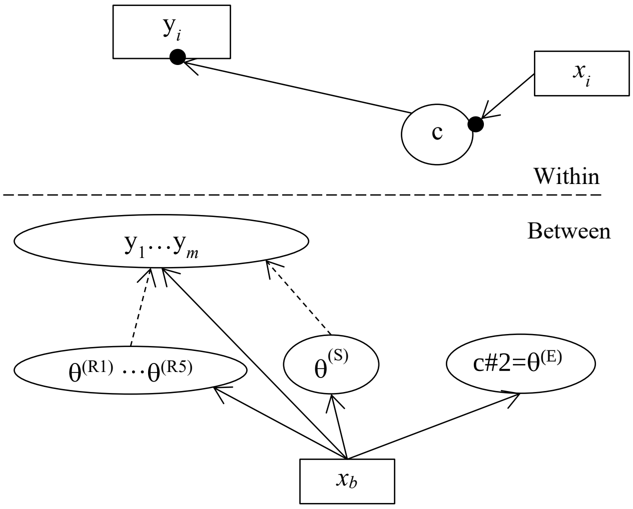
B. Factor Mixture Analysis model (FMA)

C. Faking as Grade of Membership model (F-GoM)

Note. Markers are set according to: (B) most likely class in FMA model; (C) most likely class in F-GoM model. Legends display observed-predicted correlations within each marker group.

Figure 4.

Application: Diagram of the F-GoM Model with Within-Level Covariates x_i and Between-Level Covariate x_b

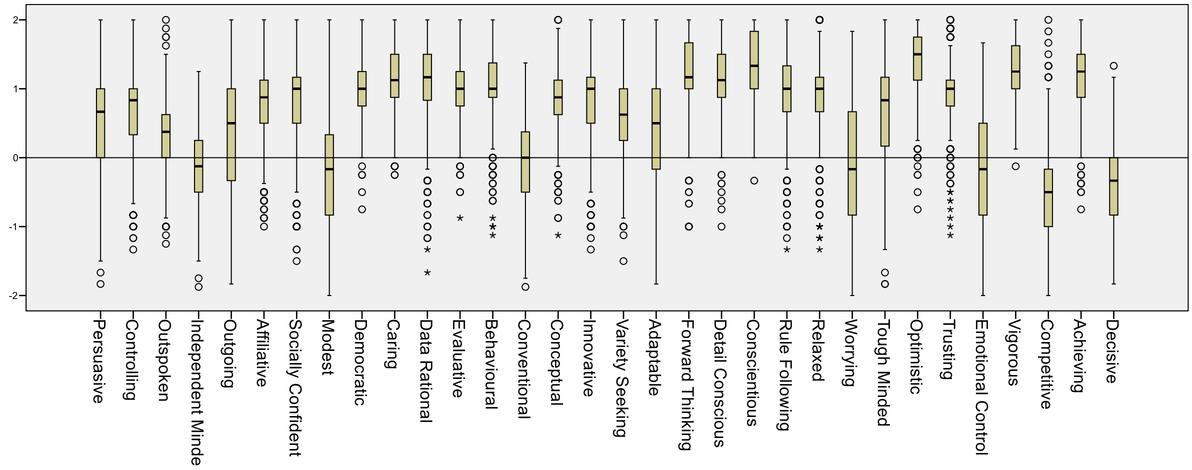


Note. c (1='real', 2='ideal')

Figure 5.

Application: Distributions of Observed OPQ32n Scale Scores (Quartiles and Outliers)

A. All responses



B. Responses by most likely class in F-GOM model with equal within-level intercepts ($\alpha_i=\alpha$)

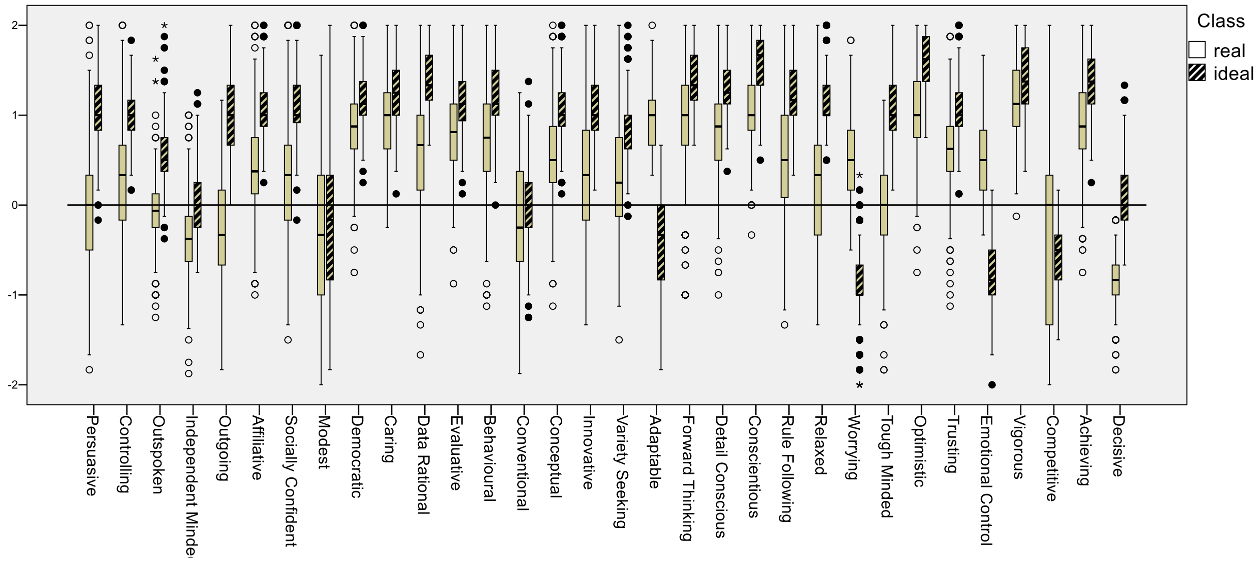


Figure 6.

Application: Scree Plot for Observed OPQ32n Scores

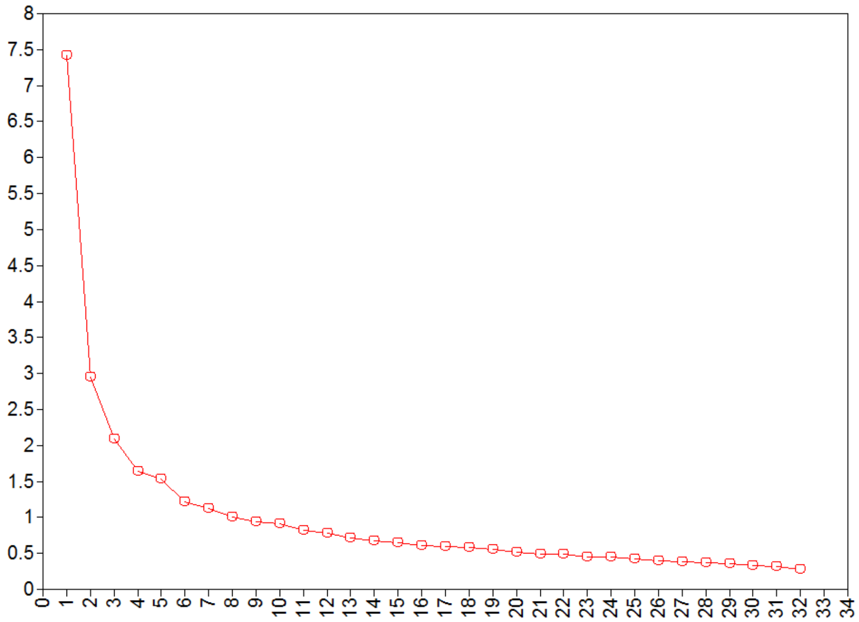
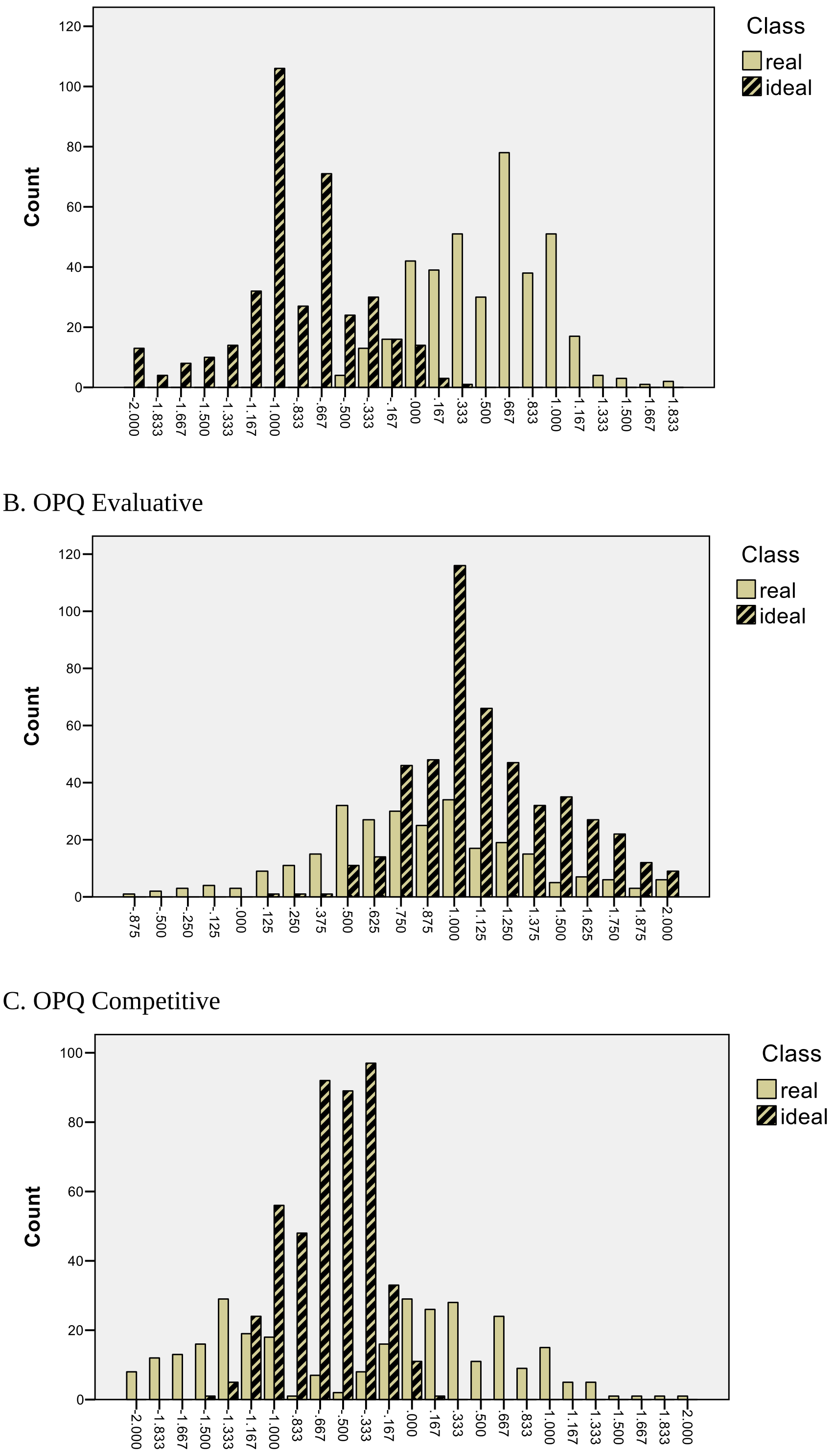


Figure 7.

Application: Frequencies of Likely 'real' and 'ideal' Responses by Observed OPQ Score Value in the F-GOM Model With Equal Within-Level Intercepts ($\alpha_i=\alpha$)

A. OPQ Worrying



B. OPQ Evaluative

C. OPQ Competitive