Can faking be measured with dedicated validity scales?

Within Subject Trifactor Mixture Modeling applied to BIDR responses

Nigel Guenole
Goldsmiths, University of London

Anna Brown
University of Kent

Velvetina Lim
University College London

This manuscript is accepted for publication at *Assessment*. Correspondence regarding it can be sent to n.guenole@gold.ac.uk.

Abstract

A sample of 516 participants responded to the Balanced Inventory of Desirable Responding (BIDR) under answer honest and instructed faking conditions in a within-subjects design. We analyse these data with a novel application of trifactor modeling that models the two substantive factors measured by the BIDR – Self-Deceptive Enhancement (SDE) and Impression Management (IM), condition-related common factors and item specific factors. The model permits examination of invariance and change within subjects across conditions. Participants were able to significantly increase their SDE and IM in the instructed faking condition relative to the honest response condition. Mixture modeling confirmed the existence of a theoretical two-class solution comprised of approximately two thirds of 'compliers' and one third of 'non-compliers'. Factor scores had good determinacy and correlations with observed scores were near unity for continuous scoring, supporting observed score interpretations of BIDR scales in high stakes settings. Correlations were somewhat lower for the dichotomous scoring protocol. Overall, results show that the BIDR scales function similarly as measures of socially desirable functioning in low and high stakes conditions. We discuss conditions under which we expect these results will and will not generalise to other validity scales.

Keywords: BIDR, validity scales, socially desirable responding, impression management, trifactor models, factor mixture models, measurement invariance, instructed faking.

Can faking be measured with dedicated validity scales?
Applying within subject trifactor mixture modeling to BIDR responses

Faking in questionnaire responses is considered such problem that remarkable levels of effort are devoted to addressing it. Validity scales are among the most commons methods for detecting dissimulation on self-report questionnaires, despite some controversy regarding the extent to which they can detect dissimulation and be used to correct for it (McGrath, Mitchell, Kim, & Hough, 2010; Morey, 2012; Rohling et al., 2011). The Marlowe Crowne social desirability scales (MCSDS) for instance, which are based on the Minnesota Multiphasic Personality Questionnaire (MMPI), differentiate attribution, or claiming desirable characteristics, from denial, or disclaiming undesirable characteristics (Millham, 1974). Leite & Beretvas (2005) reviewed the ways that the MCSDS are used in practice. They are commonly used in three different ways. First, substantive trait scores are correlated with social desirability scale scores to see if the correlations are high, indicating socially desirable responding impacted the assessment process. Second, factor analysis is sometimes used to check if the construct of interest and social desirability are empirically distinct. Finally, the substantive assessment scores for respondents with high social desirability scores are sometimes ruled invalid.

Another well-known scale for detecting social desirability that is used in similar ways to the MCSDS is the Balanced Inventory of Desirable Responding (BIDR: Paulhus, 1984; Paulhus, Bruce, & Trapnell, 1995). In early writings Paulhus traced the history of researchers who have distinguished two forms of socially desirable responding, self-deceptive enhancement (SDE) and impression management (IM), in questionnaire responses. Self-deceptive enhancement refers to when a respondent really believes their inflated responses, while impression management occurs when a respondent consciously inflates responses (Paulhus, 1984). According to Gignac (2013), the former behaviors are observable only to the self while the latter are observable to others. Paulhus identified scales that had historically been used as markers of each form of dissimulation, and noted that factor analysts who initially discovered these tendencies referred to them as alpha and beta (e.g. Block, 1965; Wiggins, 1964). Paulus (1984) factor analyzed the scale scores for questionnaires thought to be markers of alpha and beta to create the BIDR. Readers may refer to Paulhus's original work, or to Leite & Beretvas (2005) for a brief history of the origins of the BIDR, including its psychodynamic roots that were subsequently dropped.

The BIDR scales have now been extensively evaluated, including testing under instructions to fake and not to fake. This work has primarily occurred at the observed variable level with classical test theory rather than with latent variable models. Despite an abundance of research articles on the BIDR, the number of articles that explicitly focus on the measurement invariance of the BIDR under low and high stakes conditions is limited to a few papers at most. Yet the factorial invariance of the BIDR across settings where people are not expected to dissimulate and where they might is critical to the validity of the measure. For the scales to be useful in measuring the extent of socially desirable responding, the measurement parameters of the confirmatory factor models (i.e., loadings and intercepts or thresholds) ought to be the same between honest and faking conditions with only the population heterogeneity parameters (i.e., latent factor means and variances) changing between the conditions. That is, to measure the extent of SDE and IM, the scales should 'work' equally well under low and high stakes conditions. This is even more important for social desirability scales than it is for other constructs. After all, it is social desirability scales that researchers purport to be measures of faking. If they functioned differently when people faked, it would be analogous to a ruler functioning differently when measuring shorter as opposed to longer distances. Whether or not the measurement parameters vary as a function of instruction conditions is an open question, and we explore this as a research question rather than a directional hypothesis. There have been few comprehensive investigations of measurement invariance for validity scales across honest and faking conditions to our knowledge.

In fact, we found just one instance where a study examined measurement invariance for the full BIDR across low and high stakes conditions (Li & Reb, 2009). This study used a multiple group approach with a within-subjects design, which violates the assumption that the groups are independent populations and renders the conclusion challenging to interpret. Instead, a single group longitudinal invariance model should be applied with within subject designs, as discussed by authors including Chan (1998), Marsh & Grayson (1994) and Liu et al. (2017). A second study using a between groups design found support for measurement invariance for the impression management scale of the BIDR. This study was designed to examine the effect of anonymity versus confidentiality assurances (Miller & Ruggs, 2014). However, it omitted the self-deception scale. The main objective of this article is to appropriately examine the measurement invariance of the BIDR across honest and faked responses using a within subjects instructed faking design. In particular, we will determine whether any changes in item means and covariances between conditions affects measurement parameters, population heterogeneity patterns, or both.

Before invariance over experimental conditions can be examined, one needs to obtain a well-fitting measurement model for the BIDR. Few multiple factor psychological measures are truly orthogonal, and it is likely that the BIDR factors of SDE and IM are correlated. In situations where there are three or more substantive factors, potential measurement models would include a higher order factor model, a correlated factors model, or hierarchical models.

Higher order models. Higher order models model the relationships between factors with a higher order factor that explains the variance in lower-level factors. As Paulhus (1984) originally proposed a two-factor model of socially desirable responding, a higher order factor model is not identified without modelling additional variables or including model constraints beyond typical identification constraints when three or more latent variable indicators are available. This leaves the correlated factors model and hierarchical factor models (Holzinger & Swineford, 1937; Markon, 2019; Reise, 2012) as plausible models.

Correlated factor models. Correlated factor models represent the relationships between measured factors with non-zero correlations. The correlated factors model has not proved to be the best fitting model for the BIDR in past research, which has raised concerns about the BIDR structure (e.g., Gignac, 2013; Leite & Beretvas, 2005). We expect that this is because of at least two reasons. First, we expect that a general latent factor representing non-uniform response biases might be necessary (Brown, Inceoglu, & Lin, 2017). Second, we expect that unique item factors might be required, albeit that item specific factors are not identifiable in typical self-report single occasion responses (e.g., Rao, 1955).

Hierarchical models. Hierarchical models assume that all factors influence items directly, as opposed to indirectly via subordinate latent factors in the case of higher order factor models. In the bifactor model, one form of hierarchical models, a general factor is fitted along with an orthogonal subset of homogenous group factors accounting for variance not explainable by the general factor. The group factors can themselves be correlated or uncorrelated (Holzinger & Swineford, 1937; Markon, 2019; Reise, 2012).

Caution has been suggested in interpretation of bifactor models as revealing substantive group factors as opposed to its more conventional use for examining the extent to which a measure provides a unidimensional score. Reasons include the difficulty of interpreting the general factor as a causal factor; the tendency of bifactor models to improve fit by modelling construct irrelevant variance, and the fact that good model fit does not indicate validity of the latent variables in the bifactor model (Bonifay & Cai, 2017; Bonifay, Lane, & Reise, 2017). Nonetheless, these are popular models and have been applied with BIDR data.

One application of bifactor modelling to the BIDR was presented by Gignac (2013) who tested an extensive array of models using a low stakes sample that included Paulhus & Reid's (1991) updated BIDR structure, where self-deceptive enhancement is split into self-deceptive enhancement and self-deceptive denial. Gignac compared the fit of numerous models where the data were dichotomously scored and 'continuously' scored (i.e., treated as ordinal rather than binary, in both cases he modelled the data using a diagonally weighted least squares (DWLS) estimator).

In that study, the best fitting model for the BIDR's continuous scoring was a hierarchical model that included a) a general social desirability factor; b) orthogonal specific factors for self-deceptive enhancement and impression management, and c) a method factor that modelled all items that were negatively keyed. Considering the earlier concerns about giving substantive interpretations to methods factors, readers might think twice today about the validity of a bifactor representation of the BIDR with substantive group factors. At the same time, the bifactor model is not up to the complexity of the task of modelling BIDR responses in an instructed faking design, where the same participants answer honestly and under faking instructions.

However, while the modelling task becomes more complex due to the introduction of a within subjects repeated measures design, possibilities are opened by the additional experimental condition that enable identifying different sources of variance in the response process. Namely, unique item factors are now identifiable with two measurement occasions corresponding to participant responses under each instruction condition. Moreover, the latent mean differences between experimental conditions can be identified, and if appropriate, interpreted. In this article, we capitalize on this opportunity to present a novel application of tri-factor modelling (Bauer et al., 2013), which is itself an extension of the bifactor model.

Trifactor models. Bauer et al., (2013) presented trifactor modelling in the context of multi-informant designs, suggesting that responses of multiple informants answering about a single construct could be explained by a general factor measuring the construct of interest, rater factors measuring the unique perspective of each rater source on the latent construct, and item specific factors that measure unique variance associated with each item across the informant groups.

We adapted Bauer et al.'s approach in the following ways. We modelled two **substantive BIDR factors** in each experimental condition. Hence, the model included honest SDE and faked SDE and honest IM and faked IM factors. These are the ultimate constructs of interest when one uses the BIDR, because they reflect the extent of respondents' self-

deceptive enhancement and impression management under low and high stakes assessment conditions. Different to Bauer et al., (2013), we incorporated a *mean structure* for the BIDR substantive factors, permitting the interpretation of the experimental difference in latent means resulting from the faking intervention. To this end, we imposed strict measurement invariance across conditions, and fixed the means of SDE and IM in the honest condition to zero while estimating them freely in the faking condition. The measurement invariance constraints are described in the Methods section. Next, we modelled further dependencies in item responses in each condition with two **method factors** – thus, all 40 BIDR items rated under the honest condition loaded on a "general Honest" factor, and all 40 BIDR items rated under the faking condition loaded on a "general Faking" factor. Brown et al., (2017) suggested that such general factors can be used for capturing response biases. If the factor loadings are similar across items, they can be considered uniform distortion (i.e., response styles), such as acquiescence, while if they are different across items (for example, follow the pattern of positive and negative loadings in the BIDR's balanced design) they can be considered non-uniform biases, such as socially-desirable responding. However, in the present study the latter is measured directly via the BIDR, so the general factors are intended to pick up any remaining variance due to response styles. Luckily, the BIDR's balanced design provides a unique opportunity to separately identify the substantive factors SDE and IM (with half of the items expected to load negatively), and the method factors as response styles (with all items expected to load positively). Finally, dependency that is due to the same item being answered in two instructional conditions was modelled as **item specific factors**, operationalized as correlated errors of the same item across the two conditions. This parameterization is mathematically equivalent to the item specific factors with both factor loadings set to unity and freely estimated variance, producing identical fit. This trifactor model variation is represented graphically in figure 1.

Insert figure 1 about here

While modelling method-related general factors on top of the substantive factors often cause problems with model identification (e.g., Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), the repeated measures design in this study allows identification of substantive factors by imposing strict measurement invariance across conditions, which in turn allows assigning the mean differences between BIDR items in honest and faking conditions to the latent SDE and IM mean shifts rather than to the method factors. This would

not be possible without the repeated measures design. To further facilitate this separate identification, as is recommended in the bifactor modelling literature (e.g., Reise, 2012), we set the method factors uncorrelated with the substantive factors. The method factors, however, should correlate with each other if they are to capture the same person's response styles under different instructions. Similarly, the substantive factors should correlate with each other if they are to capture the same person's standing on SDE and IM under different instructions.

With reference to the concerns with interpretation of bifactor models mentioned earlier, it is important to note that we do not make any causal interpretation of the BIDR factors that were not originally proposed by Paulhus. The hypothesised substantive factors in each condition are still impression management and self-deceptive enhancement. The trifactor model is simply a technical adaptation that allows us to interpret the impact of faking on a) the measurement properties of the self-deception and impression management scales under high stakes conditions, and b) if appropriate, as determined by the invariance of the item parameters across instruction sets, the differences in latent means because of the experimental manipulation. However, this is not necessarily to say that method factors in the trifactor model could not be given a substantive interpretation that is not based on response styles at a later point given appropriate evidence.

Trifactor mixture models. In any experimental intervention there is a chance that some participants will not follow instructions, and we anticipate that in the faking condition, which imposes a significant cognitive burden on participants, some participants do not actually fake but respond normally. Hence, the distribution of scores in the faking condition may show heterogeneity as a result a mixture of honest and faked responding. We accommodate for this scenario with an adapted trifactor mixture model, which identifies two latent classes of individuals (Clark et al., 2013), such as people who followed the faking instructions and those who did not. Within the 'compliers' class, the default trifactor model with the honest and faking conditions applies; and within the 'non-compliers' class, both repeated measures in the trifactor model are modelled as the honest condition. Kim & von der Embse (2020) described the integration of trifactor modeling and mixture modeling in the context of Bauer et al., (2013)'s original formulation of the trifactor model. Here, we extend the application of trifactor mixture modeling to the repeated measures designs to allow modeling simultaneously the dimensional structure of the BIDR across faking and non-faking while also identifying classes of individuals who complied (compliers) with the instructions and those who did not (non-compliers).

Method

Participants

Here we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Participants in this study were a) 213 professional colleagues and students in the network of the author team and professionals in the working community in the United Kingdom who consented to participate after the survey was advertised to professional networks and on the LinkedIn website, and b) 303 respondents from a panel survey company called, Cint, who were representative of British working adults. The study was approved by the ethics committee at the first author's institution. Sample (a) received the instruction to fake first, followed by the instruction to respond honestly; and this order was reversed in sample (b). The two samples were similar in some but differed in other potentially important respects. The largest demographic group was white in both cases; but the student sample was younger on average and its modal highest education was higher. We present a detailed demographic breakdown of the samples in Table 1, including age, gender, education, occupation and ethnicity.

There were significant differences in observed scale scores between the samples, presented in Table 2, that is impossible to clearly attribute to the different sampling mechanisms or the fact that the first sample was asked to answer honestly before faking rather than vice versa. Our conjecture, although impossible to prove, is that the bigger shifts in means between conditions for the student sample were because they first had the opportunity to answer the items honestly before being asked to fake. Nonetheless, in both samples, the means shifted due to the experimental conditions in the same direction. We analyze the sample overall, as each sample alone is too small to analyze with the trifactor model set up we use. Most interpretations of sample size requirement for confirmatory factor models require at least 400 participants, particularly with the MLR estimator with missing data (Savalei & Bentler, 2005; Yuan & Bentler, 2000) and a highly parameterized model.

Insert Tables 1 and 2 about here

Measures

In this paper, we analyse and report on Paulus's (1984) BIDR scales administered as part of a larger study. Paulhus discussed the use of both five-point and seven-point Likert scale variations as well as dichotomous scoring schemes. In this study we administered the

BIDR with a five-point scale ranging from ‘strongly disagree’ (1) to ‘strongly agree’ (5). In addition to demographic data questions described under the participants section above and the BIDR measures, participants completed a maladaptive personality measure, the G50 (Guenole, 2015), which is not analysed here. The exact design reported in (Guenole, Brown, & Cooper, 2018) was completed under honest and faked conditions.

Design and Procedure

Assessment stakes (low versus high) were experimentally manipulated within subjects. In the ‘low stakes’ (‘honest’) condition, respondents were instructed to answer honestly as follows: ‘You are now going to be answering questions about your personality. We would like you to answer questions in this section as honestly and as accurately as possible. Please answer truthfully, all responses are anonymous’. In the ’high stakes’ (’instructed faking’) condition, respondents were asked to respond as though they were applying for a job they really liked, with the following instructions: ‘In this section we would like you to 'fake' your answers to the questions. In other words, answer as if you want to make your best impression to get a job you really want’.

Many research studies employing instructions to fake specify a particular job type so that all respondents are faking towards the same profile (Robie, Brown, & Beaty, 2007; Wetzel, Frick, & Brown, 2021). The counter argument to using this type of instruction with heterogeneous samples is that this would likely lead to different people being better positioned to fake for the nominated job role than others, in addition to having very different motivations and attitudes towards the nominated job profile, which may lead to the lack of interest or motivation to follow the faking instruction in many participants. Moreover, we expect that heterogeneity of job roles that participants imagine should not be detrimental for the analysis of BIDR data, which reflect general rather than job-specific considerations of socially acceptable behaviours and unlikely virtues.

Analysis

All data and scripts used in this article are available at the following link: https://figshare.com/s/8b0cb6a9b046a864c42d. The measurement model fitted was a variation of the trifactor model first reported by (Bauer et al., 2013) and discussed above. We identified the unit of measurement for the BIDR scales by identifying a strong loading item in each scale (the referent) in each condition and setting its factor loading to one in each condition.

Estimation. There can sometimes be confusion when BIDR models are described as continuously *scored*, which is contrasted with dichotomous scoring conventions that were proposed by (Paulhus, 1984), and when researchers discuss whether model parameters were estimated using continuous or categorical estimators. We considered five-point Likert data in this study, because we were interested in modelling the response process and identifying various variance sources. In this section, therefore, references to continuous and categorical refer to parameter estimation methods for the five-point Likert data that we analyse, not to any BIDR scoring protocols. It is possible to model these response data as continuous or ordered categorical, using maximum likelihood with robust standard errors (MLR) or a diagonally weighted least squares (DWLS) estimator respectively, implementing the appropriate identification constraints for repeated measures measurement invariance (Liu et al., 2017). On the one hand, these data are certainly ordinal, suggesting DWLS. On the other hand, it is unlikely that response tendencies underlying the observed variables (assumed in all limited information estimators such as DWLS) are multivariate normal, particularly in the instructed faking condition.

Inspection of the item distributions for the continuous scoring protocol, presented in figure 2, reveals an increase in endorsement of extreme categories in the instructed faking condition, making the distributions of item responses heavily skewed. This likely reflects the respective skewness of the underlying response tendencies (which represent utilities or psychological values that people feel towards the items). We can see this same pattern in the observed scale totals in upper panel of figure 3. The categorical analysis will, however, treat the increased frequencies in extreme categories by widening the boundaries of these categories to preserve the normality of the response tendency, thus totally distorting the estimates for thresholds and polychoric correlations. A similar point was made by Robitzsch (2020). Using ML with continuous data shifts the assumption of multivariate normality to the observed level, but MLR is robust to non-normality in the observed variables. As we expect that that the continuous response interpretation is more representative of the actual response process due to the expected latent utility non-normality, we analyzed these data with the maximum likelihood estimator with robust standard errors in Mplus 8.6 (Muthén & Muthén, 2009).

Insert figures 2 and 3 about here

Data cleaning and missing data. Some participants failed to complete some or all of the honest or faking instruction conditions. As a result, we undertook the following data screening analyses. First, we removed the respondents who had missing data for more than 50% of questions in either the honest or the instructed faking condition. Even optimistic interpretations of missing data treatment approaches are expected to struggle with greater missing data than this. Second, we eliminated any participants who completed the entire survey (including the additional maladaptive survey items) in under 10 minutes. We deemed it implausible that a respondent could pay due attention to the questionnaire in such a short time. All pairwise samples following this data screening were greater than 95%, and missing data was handled with full information maximum likelihood (Enders & Bandalos, 2001).

Model fit. We first examined the fit of the baseline model from which all invariance tests would be conducted. We examined χ^2, the root mean square error of approximation (RMSEA: Steiger, 1998), the comparative fit index (CFI: Bentler, 1990), the Tucker Lewis index (TLI: Tucker & Lewis, 1973) and the standardized mean root square residual (SRMR: Jöreskog & Sörbom, 1989). A significant p-value for the chi-square indicates rejection of the fitted model. For the other indices, values observed are compared against established cut-offs (Hu & Bentler, 1999; Hu & Bentler, 1998). For RMSEA, .05 has been suggested as indicating close fit, and .08 for adequate fit. For CFI, a cut-off .95 is considered for good fit and .90 for adequate fit. For SRMR, a value of .08 or less is considered acceptable. In addition to looking at these global fit statistics and indices, we inspected the size, sign and significance of the parameter estimates themselves (e.g., item loadings and latent correlations), as well as the correlation residuals and modification indices (Kline, 2015).

Measurement invariance. We implemented longitudinal mean and covariance structures (LMACS: Chan, 1998) approach appropriate for the repeated measures design employed in this study. In this approach, the repeated-measures factor invariance model is identified by fixing the factor loading of a single referent item to one at each timepoint (i.e., condition), constraining the intercept of the referent item equal across conditions/times, fixing the latent factor mean of the first timepoint (condition) to zero, and freely estimating it in the other. The referent item for each factor was identified as an item that loaded strongly in each condition. To test for invariance at the item level from this baseline, we used the free-baseline method described by Stark, Chernyshenko, & Drasgow (2006).

This approach imposes additional item constraints to the free-baseline model one by one, constraining each item's loading and intercept simultaneously across conditions, and comparing fit with the free baseline that has only identification constraints in place. If the

change in fit is statistically significant, then measurement non-invariance, or differential item functioning (DIF), is detected. We adopted a p-value of .01 as our criterion for statistical significance given that we were making many comparisons in total in this sequence of analyses. Each item has three loadings, one for the substantive factor (either SDE or IM), one for the condition factor, and one for the item specific factor. We conducted item level invariance analyses on the substantive factor loading (and intercept) only as only the invariance of the constructs measured by the BIDR itself can be plausibly expected. The invariance for the condition factor is not assumed, and the invariance of the item-specific factors is given by their fixed to unity loadings. Tests were conducted item by item and we applied Satorra & Bentler's (2001) correction when calculating χ^2 between nested models.

Mixture Modeling. We estimated the mixture model variation of the trifactor model using the MLR estimator which permitted comparison of the single class trifactor model with partial measurement invariance we report with a two class trifactor mixture model. Important differences between the complier class and the non-complier class are that in the non-complier class the latent means of SDE and IM in the faking condition are zero, just as they are for these factors in the honest condition. In contrast, an increase in the latent scores is expected between the same construct in different conditions for the complier class. Second, given that there is no change for the non-complier class, the correlation between corresponding factors, SDE honest and SDE faked, IM honest and IM faked, is expected to be near one indicating no change in the participant rank ordering across conditions. In contrast, we expect correlations substantially less than one between corresponding factors across conditions in the complier class. Given our a priori anticipation of two latent classes, this application is considered a confirmatory trifactor mixture model.

To compare the fit of trifactor mixture models, for which chi-square based fit indices are not available, to the ordinary trifactor models, we used information criteria - Akaike's Information Criterion (AIC; Akaike, 1987) and the Bayesian Information Criterion (BIC; Schwarz, 1978). When alternative models are compared, the model with smallest AIC/BIC is considered best, and the AIC difference of 10 or greater with the alternative model is interpreted as 'providing no support' for the alternative model (Burnham & Anderson, 2004).

Finally, we examined entropy, which reflects the class separation. The higher the entropy the clearer the class separation, and values of .80 and above have been suggested as indicating strong class separation (Asparouhov & Muthén, 2014). We consider all of these in evaluation of the confirmatory trifactor mixture model.

Results

Global fit

Fit statistics for all models that we discuss in the following sections are presented in Table 3 to facilitate model comparison. The **baseline (unconstrained) trifactor model** included the two BIDR substantive factors in each condition, condition related common factors, and specific factors representing the same item asked across conditions. We only included constraints required to identify the model including the latent mean difference between conditions. The fit for model, estimated with the continuous MLR estimator, was as follows: χ^2 = 4807.552, df = 2953, p = <.01, RMSEA = 0.035 (90% CI: .033-.037), CFI = .848, TLI = .838, SRMR = .053. While χ^2 was significant, RMSEA and SRMR indices indicated adequate fit while incremental fit indices fall short of conventional cut-offs for good fit. The trifactor baseline model for the 80 item responses (40 BIDR items x 2 conditions) fitted significantly better than a correlated factors model with no condition factors or specific factors, for which the fit was χ^2 = 7924.564, df = 3074, p = <.01, RMSEA = .055 (90% CI: .054- .057), CFI = .603, TLI = .592, SRMR = .079. Adding a negatively keyed method factor, which Gignac (2013) reported for his best fitting model, marginally worsened the fit of the trifactor model. We concluded that the trifactor model of item responses treated as continuous was plausible.

Readers may be interested in the fit of the ordinal trifactor model, which was as follows: χ^2 = 4850.728, df = 2994, p =<.01, RMSEA = .035 (90% CI: .033-.036), CFI = .928, TLI = .924, SRMR = .057[1]. Gignac (2013)'s best fitting model to the raw response data (i.e., without dichotomization) was χ2 = 1379.21, df=680, RMSEA = .047 (90% CI: .046-.051), CFI = .840, TLI=.817, and SRMR was not reported. Our model, which includes an additional 40 items representing the instructed faking condition, fitted better than Gignac's model for honest responses only to 40 items on every fit criterion. It appears that with our non-normal response data the ordinal estimator overfits the data by 'normalizing' data that are actually non-normal. Supporting the idea that the ordinal model is overfitted, Savalei (2020) has discussed the tendency of incremental fit indices to overestimate the fit of models with ordinal data.

[1] This model when first fitted had a negative but non-significant variance for the specific factor for item 6 suggesting there was no specific variance for item 6 across conditions. Omission of this specific factor led to convergence to the admissible model that we report here.

Insert Table 3 about here.

Correlation residuals

We examined the model's residual correlations next. We considered the absolute value of residuals to see any that were above an absolute value .10, following (Maydeu-Olivares & Shi, 2017). Just 7% of the $n*(n-1)/2 = 3160$ residual correlations were above .10. Of those that were larger than .10, the median (and mean) correlation was .12 and the maximum was .19, for a correlation residual between unrelated items across conditions. Other correlation residuals were for seemingly unrelated items within conditions – both within and across scales. Given that these represent a very small proportion of correlation residuals, that their absolute values were not excessive, and these modifications were not anticipated a priori, we did not incorporate these empirically driven revisions in our measurement invariance testing that we report below.

Modification indices

In contrast to the residual correlations as indicators of model deficiencies in explaining particular inter-variable covariances, modification indices provide more direct advice on where changes to the model might be required by pointing to specific parameters. Modification indices did not, on this occasion, indicate substantive changes that would improve the fit that we could have a priori anticipated. For example, the largest modification indices all pointed toward modifications that were contrary to Paulhus's (1984) theory, such as allowing items to load on alternate factors or across conditions, or to changes that had no theoretical validity basis, such as residual correlations between seemingly non-related items.

Item parameters

We also inspected the parameters of the model paying most attention to the item loadings in terms of sign, size and significance and interpreting factor correlations in the same way. In Paulhus's (1984) original specification, the first item of the self-deception scale is positively phrased, the second is negatively phrased, and the remainder continue alternating sign in this manner. The first impression management scale item, in contrast, is hypothesized to be negative, the second is positive, and the remainder continue to alternate in this pattern. Inspection of the model estimated loadings on the substantive SDE and IM factors indicated that this pattern held for all except the seventh item of the self-deception scale, which should have been positive but was in fact weakly negative. We reserve

presentation of the substantive factor loadings until the measurement equivalence section where we present the equated loadings.

General factor loadings

Brown, Inceoglu, & Yin (2017) offered suggestions about how to interpret methods factors such as the general factors in this paper. They suggested such general factors can be considered random additive effects of bias, for instance, that have been incorporated in the modelling of the response process (Böckenholt, 2012). The nature of the distortion can be uniform (e.g. acquiescence and leniency factors discussed by Maydeu-Olivares & Coffman (2006) or non-uniform (e.g. the ideal employee factor described by Klehe et al. (2012)). Non-uniform distortion would be implied by non-equal factor loadings across items, while the uniform forms of distortion would be suggested if the loadings on the common factor were equal across items.

The standardized loadings for the general factor in each condition for the tri-factor model are presented in Table 4. A first observation is that the range of these loadings in each condition is narrow, from -.10 to .45. Variation is apparent in the loadings from inspection, and a formal test of the difference between the models where these loadings were freely estimated and a model where they were constrained to be equal significantly worsened fit. However, the loadings do not follow the pattern of an ideal employee factor, because both the desirable and undesirable items measuring the SED scale have positive factor loadings. This pattern is less pronounced for the IM items, where the desirable items have mostly near-zero loadings and the undesirable items have mostly positive loadings. The near-zero loadings for the desirable IM items provide reassurance that the general factors do not capture the 'faking' variance because if this were the case, the desirable IM items would be affected most as ultimate indicators of impression management behaviour. This is good news because in the trifactor model, the **change** in substantive factors SDE and IM is expected to capture the faking effect. It seems likely that because there is restricted variability of loadings, the general factor in both conditions is a mix of content-dependent acquiescence and, despite that we screened fast response times, inattentive responding. We return to ways that the inattentive responding might be eliminated in future research in our discussion.

Insert Table 4 about here

Measurement invariance for substantive factors

With the baseline model established we proceeded with item level invariance tests that simultaneously examined loadings and intercepts. These results are presented in Table 5. These results indicate partial invariance, because while for the majority of items the p-values are below .01, there are three items for self-deceptive enhancement and three for impression management that showed significant χ^2 difference tests. Six items might seem like a large number, but it should be considered as a proportion of the total of 40 BIDR items (15%). We conducted a series of follow up measurement invariance tests to examine whether differences on the slope or intercept gave rise to the significant combined tests reported in table 5. We first conducted a loading invariance test, and only if the metric (loading) invariance test was non-significant, conducted a metric (intercept) invariance test. Following these additional tests for each of the items that were initially identified as non-invariant, we attempted a qualitative examination of causes for the difference in item functioning. It is important to note that there may not be any obvious reason for the measurement non-invariance, and hence, we cautiously offer here potential reasons. If the BIDR scales were to be reviewed, we would recommend expert panel reviews that consider all items but focus on the items that showed non-invariance items as a starting point.

For the SDE scale, the first item to exhibit non-invariance was item #1, "My first impressions of people usually turn out to be right". The loading was invariant but there was non-invariance on the intercept, with the expected item score for a person scoring zero on SDE lower in the faking condition. One possible reason may be that agreeing with this statement is seen as arrogant and undesirable in high stakes employment situations. The next SDE item exhibiting non-invariance was #3: “I don't care to know what other people really think of me". With this item the non-invariance was due to the loading. Interestingly, item #3 went from being a good indicator of the SDE construct in the honest condition to almost unrelated to SDE in the faking condition. A possible explanation would be that under high stakes conditions, social desirability considerations do not apply to this item, and response decisions are based solely based on impression management considerations (which would result in almost universal rejection of the item). Finally, on the SDE scale, item #13 "The reason I vote is because my vote can make a difference" again showed the loading non-invariance. In contrast to item #3, however, the strength of the loading went from moderate in the honest condition to strong in the faking condition. It seems that in a high-stakes setting, the item triggered more consideration for self-deceptive enhancement than in a low stakes setting.

There were also three items that showed non-invariance on the IM scale. First, item #5, "I sometimes try to get even rather than forgive and forget" showed loading non-invariance. The loading was more strongly negatively related to the underlying IM factor in the faking condition, indicating its greater importance for impression-management considerations in a high stakes setting. In contrast, the final two items on the IM scale showed intercept non-invariance. These were item #6, “I always obey laws, even if I'm unlikely to get caught”, and item #10 "I always declare everything at customs". Both items had lower intercepts in the faking condition, indicating perhaps less enthusiasm for producing extreme scores where simply hiding illegal behaviour would do, in comparison to other items in high stakes settings. We note that these minor intercept decreases are after controlling for the very large IM score inflation from the honest to the faking condition, and that the means of these items are much higher in the faking condition.

We estimated a final **partial invariance model** for these data based on the measurement invariance results, allowing item loadings and intercepts to freely vary for 6 items where invariance analyses produced a significant decrement in fit. The fit of this model was χ^2 = 5053.594, df = 3018, p<.001, RMSEA = .036 (90% CI: .034 -. 038), CFI = .833, TLI = .826, SRMR = .057. After achieving partial invariance, the standardized latent means in the faking condition were 1.163 on the self-deception scale and 1.322 on the impression management scale. Hence, the positive latent means represent experimental effects in the positive direction, with both SDE and IM showing large increases (score inflation) in the faking condition. The standardized loadings for the substantive factors for the partial invariance model are presented in Table 6.

Insert Tables 5 and 6 about here

Factor score determinacy

We examined the factor determinacy for the substantive factors – the correlation between the factor score estimates and the latent traits they represent (Beauducel, 2011; Guttman, 1955). The closer factor determinacy coefficients are to 1, the better factor scores represent the latent factors, and cut values based on whether scores are used for research (.80: Gorsuch, 2014) or individual assessment (.90: Grice, 2001) have been proposed. The factor score determinacies for the trifactor model for the complete data pattern were all high as follows: Honest SDE .901; faked SDE .906; honest IM .958; and faked IM .956. Moreover,

for all of the numerous missing data patterns factor determinacy was similarly high. These values meet even the strictest recommendations for individual assessment discussed by Grice.

Mixture modeling

When we examine the BIDR observed score distribution plots for continuous responses in the faking condition we see a clear bimodal distribution, suggesting a mixture of subpopulations might account better for the response patterns in that condition. However, this is not the case in the honest condition. A plausible explanation is that under the faking condition, not all participants follow the instruction and think of ideal responses; some simply respond in the normal fashion (honestly), which is less cognitively demanding. To test this hypothesis, we allowed two latent classes in the final equated trifactor model as described earlier. The first class is a class of "compliers", the second class is a class of "non-compliers". In the mixture model, the factor means, variances and covariances are allowed to differ in the class on non-compliers, but only in the faking condition. All parameters related to the honest condition are the same across both classes because people in both classes complete the honest condition normally.

The **trifactor mixture model** has only 16 parameters more than the equated trifactor model, and it fits decisively better. The AIC (114852 v 114587) and the BIC (116134 v 115938) were both much lower for the two-class model. The mixture model also has very interpretable average response profiles, and the expected correlations of near 1 between respective constructs even though these were freely estimated. The model also has good class separation, indicated by an entropy value of .88. With the mixture model, the bimodal distributions of the latent variables are clearly explained as an artefact of two subpopulations present – compliers and non-compliers, and as a result, there was a higher faking effect than in the standard trifactor model, represented by the standardised mean differences of 2.235 for SDE and 2.335 for IM between honest and faking conditions. This is because 30.6% of participants actually belong to the non-compliers class, and their honest responses were dragging the overall effect of the faking instruction down in the single class model.

Factor correlations for the mixture model

Substantive factors of the BIDR, SDE and IM, were allowed to correlate with one another within conditions and across conditions. For the complier class the within condition correlation for the substantive factors was .27 for honest instructions and .79 for faking instructions. The within condition correlation for the non-complier class was also .27 for the

honest instructions and was .25 for the faking instructions. For the complier class, the across condition correlations between corresponding substantive factors was .14 for SDE and .20 for IM, while in the non-complier class these values were 1.06 for SDE and 1.03 for IM. We note that for non-compliers these correlations are expected to be unity, and while they are estimated as slightly greater than unity the confidence intervals for both correlations include one.

General factors were set to be orthogonal to substantive and specific factors but were allowed to correlate with each other across conditions. The correlation between the general factors in the honest and faking instruction was .54 for the complier class and 1.02 for the non-complier class. Once again, these are expected to be unity for non-compliers by design and while they are estimated greater than unity confidence intervals about these estimates include the expected value of one. Item specific factors were orthogonal to all other factors as well as being orthogonal to one another and are not interpreted.

Observed to latent scale correlations

We expect that most readers will use the BIDR observed scores, so a natural question might be, what is the relationship between observed BIDR subscale scores and latent representations of the same dimensions? We calculated the correlation of the estimated factor scores in the best-fitting model, **trifactor mixture model**, with the observed scores, revealing the results presented in table 7. The correlations between the estimated factor scores for SDE and IM and the respective five-point Likert scores were all in excess of .95. This indicates that for practical purposes, the five-point scoring system will produce ordering of people similar to that described by the SDE and IM substantive factors modelled in this article. On the other hand, table 7 reveals that the dichotomously scored BIDR and the estimated factor scores are somewhat different. For the SDE scale, the correlation between latent and dichotomous scores were .53 and .74 respectively for the honest and faking conditions, while the correlations between latent and dichotomous scores for the IM scale were .80 and .91 respectively. The lower correlations for the SDE scale across conditions, relative to the IM scale, are likely due to the BIDR dichotomous scoring protocols. For the SDE scale, this protocol sees reverse scored items appropriately recoded, and then coding one, two, three or four as zero, and values of five recoded as one. The IM scale, on the other hand, sees one, two, and three recoded as zero, and four and five recoded as one. Such dichotomisation of the IM scale is likely more representative of the threshold differentiating high and low stakes responses (with three bottom response options being almost exclusive markers of the low

stakes), therefore, it captures more information about the latent traits and aligns more closely with the continuous scoring in the trifactor model. The fact that the dichotomous scoring less accurately reproduces the continuous score distributions is evident in the distributions presented in figure 3.

Insert table 7 about here

Discussion

Socially desirable responding is a ubiquitous threat to validity in psychological assessment, so it is important that we have techniques available to identify when such responding occurs. Among the most widely used approaches are validity scales, and perhaps the most routinely used validity scales is the BIDR. Early efforts to validate the BIDR's purported two factor structure have produced unsatisfactory fit. In fact, Leite & Beretvas, 2005) commented with respect to the two factor structure of the BIDR that 'It seems that until the structure of responses to the MCSDS and the BIDR can be better clarified, researchers should be careful when attempting to correct scores of other scales based on SDB scores. (p152)'. In this study, we were able to tease apart the factors that were preventing adequate fit on the BIDR in earlier studies using a trifactor modelling approach and to establish partial measurement invariance for the BIDR across honest responding and instructed faking – a simulated high stakes situation. We also showed that the factor scores, estimated with high determinacy, correlated near unity with observed variable counterparts. This means that the BIDR can be used to quantify the extent of faking in an assessment.

Methodological contributions

In terms of methodological contributions, we demonstrated how Bauer et al.'s (2013) trifactor model can be extended from a multiple raters design to study change between experimental conditions for within subjects designs. The trifactor model we fitted included substantive factors capturing the extent of self-deceptive enhancement and impression management in each experimental condition, method factors capturing response styles in each condition, and item specific factors, and it enables identification of condition related change on the substantive factors due to the response instructions. Identification of the trifactor model was permitted by having two instructional conditions – namely, the ability to identify specific item variance in repeated administration designs. The remaining (common)

item variance was partitioned into the substantive effects with their mean shift across conditions, and the effect of response style (which turned out to be mostly acquiescence).

We also demonstrated how mixture modelling can be applied to the trifactor model (or to any factor model suitable for the task in hand) to account for heterogeneity in the distribution of observed scores in the faking condition. In our experience, this is not a rare event when research participants do not fully comply with experimental instructions that are cognitively taxing. In such cases, it is possible to identify the latent class of non-compliers in fully confirmatory fashion, by constraining their model parameters in the faking condition to be equal to the honest condition. This approach allows estimating the faking effect more accurately, by disallowing the non-compliers to influence the result. In our samples, non-compliance was estimated at 30%, which would have a strong influence on both the model fit and the substantive results if not controlled.

While we operationalise item specific factors in our models as correlated residuals, it is also possible to estimate these effects as latent variables that are orthogonal to each other and all other factors in the model, which achieves identical results. Whereas the correlated residuals approach is simpler syntactically, the latent variables approach might be preferred if covariates are expected to predict the item specific factors, or if the item specific factors themselves are to be used in prediction of other variables. In our online materials we present the baseline trifactor model estimated with both correlated residuals and with specific factors ways, achieving identical fit.

Practical implications of these results

To inform use of the BIDR in applied settings, we report correlations of the factor scores from the best-fitting model, a trifactor mixture model, with their observed variable counterparts in each experimental condition. We show that the five-point scoring scheme yields scores that correlate highly (above .95) with their trifactor model counterparts, while the dichotomous scoring departs from them substantially. These results give some confidence regarding the interpretation of the observed BIDR scores derived from the five-point scoring schemes as representative of the same factors in the latent variable model. Because the latter were shown to be mostly measurement invariant across low and high stakes instructions, and sensitive to these instructions in term of the mean shift, the observed five-point sum scores will possess similar properties and can be recommended for use in practice. The dichotomous BIDR scores, however, are not supported by the same evidence and need further investigation with respect to their construct validity.

On the broader question of using validity scales such as BIDR in practice several points are important to note. The first is that these results give some confidence regarding the use of the BIDR summated scores or the factor scores estimated and saved from respective latent variable models as a check for the extent of faking in applied assessment settings. For instance, at the individual level, observed scores that exceed the recommended cut-offs might prompt particular care to examine consistency between how individuals describe themselves in response to different assessment methods as well as self-other discrepancies on constructs assessed with multiple methods. Where non-invariant items are identified, it would be best to remove them from the sum score. At the group level, the score distributions and the means can serve as good indicators of the extent of self-deceptive enhancement / impression management in the population of test takers in the current assessment context.

Second, while we provide support for the use of the observed BIDR scores as measures of faking, this does not imply that they could or should be used to "correct" any substantive assessment scores. Many authors have warned against such "corrections" using manifest scores from validity scales, since the observed scores carry at least some "substance" – that is, variance due to stable personality attributes such as neuroticism and conscientiousness (Ones, Viswesvaran, & Reiss, 1996). Indeed, results of this study show that the five-point summated BIDR scores in the faking condition correlate weakly but significantly (.17 for both IM and SDE) with the same scores in the honest condition. Therefore using these scores to partial out the "faking" variance will lead to removing some variance related to stable personality characteristics.

Expected generality of these results

A broader question is whether this result will generalise to other validity scales, and what conditions must be met for a scale to be valid in measuring faking. From a psychometric perspective, we must recognise that the BIDR scales aim to assess psychological constructs and they do so with reasonable measurement properties (e.g., unidimensionality, reliability). For the results reported here to hold for other validity scales, it is likely important that the scales also target psychological constructs with scales that have strong psychometric properties. The MCSDS for instance, which is well grounded theoretically, aims to measure a homogenous construct, behaviours with low base rate probabilities and high social desirability (Lambert, Arbuckle, & Holden, 2016). The MCSDS items are most similar to the BIDR Impression Management items, and the construct captured by them in a high-stakes situation is likely similar to that of IM (see also Uziel, 2010). We anticipate the results of this study would likely hold for this scale. The case for validity scales that are not targeted to

capture self-presentation behaviours is not so clear. For instance, the ‘Cannot say’ scale of the Minnesota Multiphasic Personality Inventory, which is a count of omitted items, is likely to measure a response style and not a purposeful behaviour, making it difficult or even impossible to generalise these results. Assuming items to measure purposeful self-presentation behaviours, for these results to generalise it is further important the observed variable rating scales directly reflect the modelled responses. We have seen here, for instance, that as the scoring procedure is coarsened the correlations between the model-based factor scores and the observed variable counterparts depart substantially from unity.

Limitations and future directions

One potential limitation is that we combined a non-probability student sample and a second non-probability convenience sample, and there were differences in the samples on demographic background variables and on observed scale scores. While non-probability samples are common, there may be some concern due to the combined student and working adult samples. Panel respondents of working adults can be different to non-panel respondents in unknown ways that attenuate relationships between variables, it is thought that this is sometimes due to being experienced experiment participants (Chandler & Shapiro, 2016). This seems not to have happened in the current study given that the largest experimental effects, at the observed level, were for the student sample which was not a panel sample. We expect that the ability to see the items once and answer honestly improved the ability to fake. However, fitting this model to a more homogenous sample of working adults should be a future direction to check the generalizability of the conclusions about the degree to which mean levels change due to the faking instruction, and to investigate the fit of the repeated measures trifactor model.

Other limitations to this study include that we collected data across two samples for counterbalancing, and it would have been desirable to randomly assign participants to the respond honest first versus faking first conditions. Nonetheless, aside from differences in mean scores on scales across the sub-samples, the data were able to be modelled in ways that suggest the results are generalizable. Some might also feel that a professional sample from a single organization might be preferrable to a potentially heterogenous community sample. Yet heterogeneous samples are characteristic of many assessment settings such as high stakes job selection.

In this study, our experimental condition factors appeared to be a mix of acquiescence and inattentive responding. This is even though we screened out those respondents who

completed in times we deemed too fast to represent diligent responding. One future approach might be to try a direct instruction question that is not scored, such as ‘choose 5 for this question’, and eliminate any respondents who fail this check. This might offer a more concrete indicator of inattentive responding. Lastly, our study used an instructed faking design, and it is not certain that people will fake in precisely the same in an induced faking situation as they would in a real high-stakes assessment context. It is difficult to overcome this limitation, however, outside of the context of an instructed faking design such as that we adopt here.

Future directions could include generalizing the adapted trifactor approach to study multiple group models to, for instance, compare invariance across populations of interest. Another major direction is exploring external validities of the method factors and the substantive factors by adding covariates to the trifactor model. This line of investigation might also link specific factors to covariates to interpret their meaning (rather than viewing specific factors as technical devices to allow accurate estimation of other factors). It would also be worthwhile to conduct simulation studies to examine the performance of the within-subjects trifactor model, including examining the relative performance of the correlated residual and specific factor implementations. Researchers may also wish to examine the properties of factor scores for the within-subjects trifactor model, as has recently been reported for the original multiple informant trifactor formulation (Curran, Georgeson, Bauer, & Hussong, 2021).

References

Akaike, H. (1987). Factor Analysis and AIC. In *Springer Series in Statistics*. *Springer Series in Statistics* (pp. 371–386). doi:10.1007/978-1-4612-1694-0_29

Asparouhov, T., & Muthén, B. (2014). Variable-specific entropy contribution. *Recuperado de Http://Www. Statmodel. Com/Download/UnivariateEntropy. Pdf*. Retrieved from https://www.statmodel.com/download/UnivariateEntropy.pdf

Bauer, D. J., Howard, A. L., Baldasaro, R. E., Curran, P. J., Hussong, A. M., Chassin, L., & Zucker, R. A. (2013). A trifactor model for integrating ratings across multiple informants. *Psychological Methods*, *18*(4), 475–493. doi:10.1037/a0032475

Beauducel, A. (2011). Indeterminacy of Factor Score Estimates In Slightly Misspecified Confirmatory Factor Models. *Journal of Modern Applied Statistical Methods: JMASM*, *10*(2), 16. doi:10.22237/jmasm/1320120900

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. doi:10.1037/0033-2909.107.2.238

Block, J. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI*. *142*. Retrieved from https://psycnet.apa.org/fulltext/1966-02934-000.pdf

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

Bonifay, W., Lane, S. P., & Reise, S. P. (2017). Three Concerns With Applying a Bifactor Model as a Structure of Psychopathology. *Clinical Psychological Science*, *5*(1), 184–186. doi:10.1177/2167702616657069

Brown, A., Inceoglu, I., & Lin, Y. (2017). Preventing Rater Biases in 360-Degree Feedback by Forcing Choice. *Organizational Research Methods*, Vol. 20, pp. 121–148. doi:10.1177/1094428116668036

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. doi:10.1177/0049124104268644

Chan, D. (1998). The Conceptualization and Analysis of Change Over Time: An Integrative Approach Incorporating Longitudinal Mean and Covariance Structures Analysis (LMACS) and Multiple Indicator Latent Growth Modeling (MLGM). *Organizational Research Methods*, *1*(4), 421–483. doi:10.1177/109442819814004

Chandler, J., & Shapiro, D. (2016). Conducting Clinical Research Using Crowdsourced Convenience Samples. *Annual Review of Clinical Psychology*, *12*, 53–81. doi:10.1146/annurev-clinpsy-021815-093623

Clark, S. L., Muthén, B., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and Strategies for Factor Mixture Analysis: An Example Concerning the Structure Underlying Psychological Disorders. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4). doi:10.1080/10705511.2013.824786

Curran, P. J., Georgeson, A. R., Bauer, D. J., & Hussong, A. M. (2021). Psychometric Models for Scoring Multiple Reporter Assessments: Applications to Integrative Data Analysis in Prevention Science and Beyond. *International Journal of Behavioral Development*, *45*(1), 40–50. doi:10.1177/0165025419896620

Enders, C. K., & Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. doi:10.1207/S15328007SEM0803_5

Gignac, G. E. (2013). Modeling the balanced inventory of desirable responding: evidence in favor of a revised model of socially desirable responding. *Journal of Personality Assessment*, *95*(6), 645–656. doi:10.1080/00223891.2013.816717

Gorsuch, R. L. (2014). *Factor Analysis: Classic Edition*. Retrieved from https://play.google.com/store/books/details?id=LDecBQAAQBAJ

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*(4), 430–450. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/11778682

Guenole, N. (2015). The Hierarchical Structure of Work-Related Maladaptive Personality Traits. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment*, *31*(2), 83–90. doi:10.1027/1015-5759/a000209

Guenole, N., Brown, A. A., & Cooper, A. J. (2018). Forced-Choice Assessment of Work-Related Maladaptive Personality Traits: Preliminary Evidence From an Application of Thurstonian Item Response Modeling. *Assessment*, *25*(4), 513–526. doi:10.1177/1073191116641181

Guttman, L. (1955). THE DETERMINACY OF FACTOR SCORE MATRICES WITH IMPLICATIONS FOR FIVE OTHER BASIC PROBLEMS OF COMMON-FACTOR THEORY1. *British Journal of Statistical Psychology*, Vol. 8, pp. 65–81. doi:10.1111/j.2044-8317.1955.tb00321.x

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41–54. doi:10.1007/bf02287965

Hu, L., & Bentler, P. 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. doi:10.1080/10705519909540118

Hu, L.-T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*(4), 424–453. doi:10.1037/1082-989X.3.4.424

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A Guide to the Program and Applications*. Retrieved from https://books.google.com/books/about/LISREL_7.html?hl=&id=LaDsAAAAMAAJ

Kim, E., & von der Embse, N. (2020). Combined Approach to Multi-Informant Data Using Latent Factors and Latent Classes: Trifactor Mixture Model. *Educational and Psychological Measurement*, 0013164420973722. doi:10.1177/0013164420973722

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. doi:10.1080/08959285.2012.703733

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Retrieved from https://play.google.com/store/books/details?id=Q61ECgAAQBAJ

Lambert, C. E., Arbuckle, S. A., & Holden, R. R. (2016). The Marlowe–Crowne Social Desirability Scale outperforms the BIDR Impression Management Scale for identifying fakers. *Journal of Research in Personality*, *61*, 80–86. doi:10.1016/j.jrp.2016.02.004

Leite, W. L., & Beretvas, S. N. (2005). Validation of Scores on the Marlowe-Crowne Social Desirability Scale and the Balanced Inventory of Desirable Responding. *Educational and Psychological Measurement*, *65*(1), 140–154. doi:10.1177/0013164404267285

Li, A., & Reb, J. (2009). A Cross-Nations, Cross-Cultures, and Cross-Conditions Analysis on the Equivalence of the Balanced Inventory of Desirable Responding. *Journal of Cross-Cultural Psychology*, *40*(2), 214–233. doi:10.1177/0022022108328819

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506. doi:10.1037/met0000075

Markon, K. E. (2019). Bifactor and Hierarchical Models: Specification, Inference, and Interpretation. *Annual Review of Clinical Psychology*, *15*, 51–69. doi:10.1146/annurev-clinpsy-050718-095522

Marsh, H. W., & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *1*(4), 317–359. doi:10.1080/10705519409539984

Maydeu-Olivares, A., & Shi, D. (2017). Effect Sizes of Model Misfit in Structural Equation Models. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences*, *13*(Supplement 1), 23–30. doi:10.1027/1614-2241/a000129

McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, *136*(3), 450–470. doi:10.1037/a0019216

Miller, B. K., & Ruggs, E. N. (2014). Measurement invariance tests of the Impression Management sub-scale of the Balanced Inventory of Desirable Responding. *Personality and Individual Differences*, *63*, 36–40. doi:10.1016/j.paid.2014.01.037

Millham, J. (1974). Two components of need for approval score and their relationship to cheating following success and failure. *Journal of Research in Personality*, *8*(4), 378–392. doi:10.1016/0092-6566(74)90028-2

Morey, L. C. (2012). Detection of Response Bias in Applied Assessment: Comment on McGrath et al. (2010). *Psychological Injury and Law*, Vol. 5, pp. 153–161. doi:10.1007/s12207-012-9131-x

Muthén, L. K., & Muthén, B. O. (2009). Mplus. *Statistical Analysis with Latent Variables. User's Guide*, *7*. Retrieved from http://www.statmodel.com/virg_nov_course.shtml

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

Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of Self-Presentation Strategies on Personality Profiles and their Structure. *Personality & Social Psychology Bulletin*, *21*(2), 100–108. doi:10.1177/0146167295212001

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*(2), 307–317. doi:10.1037/0022-3514.60.2.307

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*, Vol. 88, pp. 879–903. doi:10.1037/0021-9010.88.5.879

Rao, C. R. (1955). Estimation and tests of significance in factor analysis. *Psychometrika*, *20*(2), 93–111. doi:10.1007/BF02288983

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, *47*(5), 667–696. doi:10.1080/00273171.2012.715555

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. doi:10.1007/s10869-007-9038-9

Robitzsch, A. (n.d.). *Why Ordinal Variables Can (Almost) Always be Treated as Continuous Variables: Clarifying Assumptions of Robust Continuous and Ordinal Factor Analysis Estimation Methods*. doi:10.31234/osf.io/hgz9m

Rohling, M. L., Larrabee, G. J., Greiffenstein, M. F., Ben-Porath, Y. S., Lees-Haley, P., Green, P., & Greve, K. W. (2011). [Review of *A misleading review of response bias: comment on McGrath, Mitchell, Kim, and Hough (2010)*]. *Psychological bulletin*, *137*(4), 708–712; authors reply 713-5. doi:10.1037/a0023327

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. doi:10.1007/bf02296192

Savalei, V. (2020). Improving Fit Indices in Structural Equation Modeling with Categorical Data. *Multivariate Behavioral Research*, 1–18. doi:10.1080/00273171.2020.1717922

Savalei, V., & Bentler, P. M. (2005). A Statistically Justified Pairwise ML Method for Incomplete Nonnormal Data: A Comparison With Direct ML and Pairwise ADF. *Structural Equation Modeling: A Multidisciplinary Journal*, *12*(2), 183–214. doi:10.1207/s15328007sem1202_1

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *The Journal of Applied Psychology*, *91*(6), 1292–1306. doi:10.1037/0021-9010.91.6.1292

Steiger, J. H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(4), 411–419. doi:10.1080/10705519809540115

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10. doi:10.1007/BF02291170

Uziel, L. (2010). Rethinking Social Desirability Scales: From Impression Management to Interpersonally Oriented Self-Control. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *5*(3), 243–262. doi:10.1177/1745691610369465

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

Wiggins, J. S. (1964). Convergences Among Stylistic Response Measures from Objective Personality Tests. *Educational and Psychological Measurement*, Vol. 24, pp. 551–562. doi:10.1177/001316446402400310

Yuan, K.-H., & Bentler, P. M. (2000). 5. Three Likelihood-Based Methods for Mean and Covariance Structure Analysis with Nonnormal Missing Data. *Sociological Methodology*, *30*(1), 165–200. doi:10.1111/0081-1750.00078

Table 1

Sample demographics

Sex	Sample A	Sample B	Total
Total Number	213	303	516
Male 1	95	66	161
Female 2	118	237	355
Age	Sample 1	Sample 2	Total
Total Number	211	302	513
Mean	41.18	23.70	30.89
Standard deviation	11.48	8.16	13.07
Education	Sample 1	Sample 2	Total
Total Number	213	295	508
Less than High School	3	0	3
High School	107	113	220
Master's Degree	30	44	74
Doctoral Degree	8	0	8
Professional degree (JD, MD)	3	1	4
Bachelor's Degree	62	137	199
Ethnicity	Sample 1	Sample 2	Total
Total Number	212	302	514
White/Caucasian	186	187	373
African American	3	5	8
Hispanic	0	2	2
Asian	11	63	74
Native American	0	0	0
Pacific Islander	2	0	2
Other	7	45	52
Occupation	Sample 1	Sample 2	Total
Total Number	212	299	511
Full-time Students	10	244	254
Architecture and Engineering	4	1	5
Arts, Design, Entertainment, Sports and Media	6	3	9
Building and Grounds Cleaning and Maintenance	3	0	3
Business and Financial Operations	13	12	25
Community and Social Service	1	2	3
Computers and Mathematics	13	1	14
Construction and Extraction	8	0	8
Education, Training and Library	17	12	29
Farming, Fishing and Forestry	2	1	3
Food Preparation and Serving Related	9	0	9
Healthcare practitioners and Technicians	7	2	9
Healthcare support	7	1	8

Installation, Maintenance and Repair	2	0	2
Legal	8	1	9
Life, Physical and Social Science	3	6	9
Military	2	0	2
Office and Administrative Support	47	7	54
Personal care and service	5	0	5
Production	8	0	8
Protective Service	3	0	3
Sales and Related	22	6	28
Transport and Material Moving	12	0	12

Table 2

Observed score scale means and standard deviations

	Continuous						Dichotomous					
	Sample A		Sample B		Combined		Sample A		Sample B		Combined	
BIDR scales	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
SDE_honest	3.15	.46	2.98	.39	3.05	.43	.13	.17	.07	.10	.10	.14
SDE_faking	3.34	.49	3.79	.52	3.61	.56	.17	.20	.30	.26	.25	.25
IM_honest	3.01	.53	2.86	.51	2.92	.52	.38	.21	.35	.19	.36	.20
IM_faking	3.31	.64	3.95	.69	3.68	.74	.47	.25	.73	.26	.62	.29

Note. BIDR = Balanced inventory of desirable responding; SDE = Self-deceptive enhancement; IM= Impression Management; M = Mean; SD = standard deviation. Continuous refers to the sample statistics when the continuous scoring protocol for the BIDR is followed; Dichotomous refers to the descriptive statistics when the dichotomous scoring protocol of the BIDR is followed.

Table 3

Fit statistics for trifactor models

Model	χ^2	df	RMSEA	CFI	TLI	SRMR
Correlated factors	7924.56	3074	.06	.60	.59	.08
Trifactor continuous	4807.55	2953	.04	.85	.84	.05
Trifactor ordinal	4850.73	2994	.04	.93	.92	.06
Trifactor partial invariance	5053.59	3018	.04	.83	.83	.06

Note. The trifactor continuous model and the trifactor ordinal model used a continuous maximum likelihood estimator with robust standard errors (MLR) and a diagonally weighted least squares (DWLS) estimator respectively. The fit reported for the trifactor partial invariance model is the model where measurement invariance constraints are added to the trifactor continuous model. χ^2 = chi-square; DF= degrees of freedom; RMSEA = root mean square of approximation; CFI = Comparative fit index; TLI=Tucker Lewis Index; SRMR = standardized root mean square residual.

Table 4

Standardized factor loadings for experimental condition (method) factors

General honest				General faking			
Item	Loading	Item	Loading	Item	Loading	Item	Loading
HS1: First impressions	.21	HI1: Tell lies	.32	FS1	.31	FI1	.36
HS2: Bad habits	.27	HI2: Cover mistakes	.21	FS2	.43	FI2	.03
HS3: Don’t care others think	.44	HI3: Taken advantage	.44	FS3	.40	FI3	.45
HS4: Honest with self	.41	HI4: Never swear	.09	FS4	.43	FI4	-.10
HS5: Know why	.25	HI5: Try to get even	.37	FS5	.22	FI5	.38
HS6: Emotions biased thinking	.24	HI6: Always obey laws	-.01	FS6	.39	FI6	-.02
HS7: Made up mind	.14	HI7: Said something bad	.21	FS7	.12	FI7	.33
HS8: Safe Driver	.16	HI8: Hear people talking privately	.21	FS8	.21	FI8	.08
HS9: Control of fate	.26	HI9: Too much change	.30	FS9	.25	FI9	.45
HS10: Stop disturbing thought	.23	HI10: Declare everything at customs	.04	FS10	.36	FI10	.06
HS11: Regret decisions	.30	HI11: Stole things when young	.22	FS11	.15	FI11	.21
HS12: Can't make up mind	.20	HI12: Never littered	.07	FS12	.37	FI12	-.02
HS13: Reason I vote	.13	HI13: Sometimes speed driving	.18	FS13	.18	FI13	.28
HS14: Parents not fair	.35	HI14: Never read sexy books	-.01	FS14	.38	FI14	.02
HS15: Completely rational	.19	HI15: Done things I don’t tell	.24	FS15	.13	FI15	.36
HS16: Rarely appreciate criticism	.25	HI16: Never take things	.05	FS16	.38	FI16	.12
HS17: Confident of judgements	.25	HI17: Sick leave when not sick	.20	FS17	.19	FI17	.34
HS18: Ability as a lover	.27	HI18: Never damaged library book	.08	FS18	.34	FI18	.10
HS19: Alright if people dislike me	.41	HI19: Have awful habits	.36	FS19	.29	FI19	.33
HS20: Don’t know reasons	.30	HI20: Don't gossip	.27	FS20	.33	FI20	.13

Note. HS is Honest Self Deceptive Enhancement. HI is Honest Impression Management. FS is faking Self Deceptive Enhancement. FI is Faking Impression Management. Key words in columns may be used to match items to the original BIDR items. Key words are presented for honest

condition only as they are the same for the faking condition. General honest refers to the loadings of items in on the general factor in the honest condition. General faking refers to loadings of items on the general factor in the faking condition.

Table 5

Item level measurement invariance results

Self-deceptive enhancement				Impression management			
Item	χ^2	χ^2 diff	p	Item	χ^2	χ^2 diff	p
SD20: Don’t know reasons	4807.55			IM1: Tell lies	4807.55		
SD1: First impressions*	4835.89	23.15	.00	IM 2: Cover mistakes	4809.62	2.07	.36
SD2: Bad habits	4810.23	2.68	.26	IM 3: Taken advantage	4813.69	6.57	.04
SD3: Don't care others think**	4831.84	19.94	.00	IM 4: Never swear	4813.47	5.92	.05
SD4: Honest with self	4812.79	4.83	.09	IM5: Try to get even**	4817.86	10.3	.01
SD5: Know why	4812.11	4.56	.10	IM6: Always obey laws*	4819.47	9.48	.01
SD6: Biased thinking	4813.61	5.73	.06	IM7: Said something bad	4809.59	1.85	.40
SD7: Made up mind	4808.73	1.61	.45	IM8: Hear people talking privately	4810.68	3.12	.21
SD8: Safe driver	4808.12	.57	.75	IM9: Too much change	4810.52	2.92	.23
SD9: Control of fate	4810.6	3.05	.22	IM10: Declare everything customs*	4822.99	10.63	.01
SD10: Stop disturbing thought	4809.2	1.64	.44	IM11: Stole things when young	4814.59	5.54	.06
SD11: Regret decisions	4808.52	1.23	.54	IM12: Never littered	4808.93	1.6	.45
SD12: Can't make up mind	4812.37	4.82	.09	IM13: Sometimes speed driving	4814.22	6.27	.04
SD13: Reason I vote**	4820.39	11.73	.00	IM14: Never read sexy books	4812.43	4.69	.10
SD14: Parents not fair	4808.48	1.41	.50	IM15: Done things I don’t tell	4810.33	2.78	.25
SD15: Completely rational	4810.01	2.46	.29	IM16: Never take things	4818.72	7.4	.03
SD16: Rarely appreciate criticism	4813.76	6.21	.05	IM17: Sick leave when not sick	4810.01	2.33	.31
SD17: Confident of judgements	4818.22	8.59	.01	IM18: Never damaged library book	4821.17	9.07	.01
SD18: Ability as a lover	4810.08	2.41	.30	IM19: Have awful habits	4809.56	2.15	.34
SD19: Alright if people dislike me	4810.58	3.12	.21	IM20: Don't gossip	4810.06	2.80	.25

Note. SD is Self Deceptive Enhancement. IM is Impression Management. χ^2 is the chi-square model fit; χ^2 diff is the difference in χ^2 between the baseline model where identification constraints only are imposed and the model where an item is constrained to have the same loading and intercept across honest and faking conditions. SD is delf deceptive enhancement. IM is Impression Management. SD20 is the last item of the SD scale which was selected as the anchor item because it had strong loadings across conditions whereas item SD1 did not. The first item of the IM scale did have strong loadings in both conditions and was therefore selected as the referent for this scale. There is no χ^2 diff for the first anchor item listed in each scale because that item provides the model fit from which all other differences are calculated. * Item is not invariant with regard to intercept; ** item is not invariant with regard to factor loading. Degrees of freedom for the baseline chi-square in row 1 of the table is 2953 while degrees of freedom for the nested models is 2955. WLSMV correction value was 1.13. Key words in columns may be used to match items to the original BIDR items.

Table 6.

Standardized factor loadings for substantive BIDR factors in honest and faking conditions

Self-Deceptive Enhancement			Impression Management		
Item	Honest	Faking	Item	Honest	Faking
SD1: First impressions	.17	.27	IM1: Tell lies	-.44	-.53
SD2: Bad habits	-.49	-.67	IM 2: Cover mistakes	.36	.39
SD3: Don't care others think**	.26	-.04	IM 3: Taken advantage	-.35	-.48
SD4: Honest with self	-.45	-.57	IM 4: Never swear	.52	.62
SD5: Know why	.35	.54	IM5: Try to get even**	-.32	-.50
SD6: Biased thinking	-.54	-.66	IM6: Always obey laws	.47	.67
SD7: Made up mind	-.15	-.19	IM7: Said something bad	-.49	-.65
SD8: Safe driver	-.13	-.15	IM8: Hear people talking privately	.47	.64
SD9: Control of fate	.39	.58	IM9: Too much change	-.33	-.50
SD10: Stop disturbing thought	-.56	-.71	IM10: Declare everything customs	.37	.54
SD11: Regret decisions	.41	.50	IM11: Stole things when young	-.35	-.57
SD12: Can't make up mind	-.50	-.66	IM12: Never littered	.41	.58
SD13: Reason I vote**	.21	.59	IM13: Sometimes speed driving	-.36	-.56
SD14: Parents not fair	-.21	-.32	IM14: Never read sexy books	.25	.37
SD15: Completely rational	.44	.65	IM15: Done things I don’t tell	-.47	-.54
SD16: Rarely appreciate criticism	-.33	-.46	IM16: Never take things	.37	.54
SD17: Confident of judgements	.44	.66	IM17: Sick leave when not sick	-.48	-.68
SD18: Ability as a lover	-.45	-.63	IM18: Never damaged library book	.31	.47
SD19: Alright if people dislike me	.19	.30	IM19: Have awful habits	-.42	-.62
SD20: Don’t know reasons	-.50	-.61	IM20: Don't gossip	.44	.56

Note. SD is Self-Deceptive Engagement. IM is Impression Management. ** Item is not invariant with regard to factor loading. For all items except those marked with ** the unstandardized loadings are equated, loadings in the table differ across conditions only due to standardization. For unstandardized loadings please see the online materials.

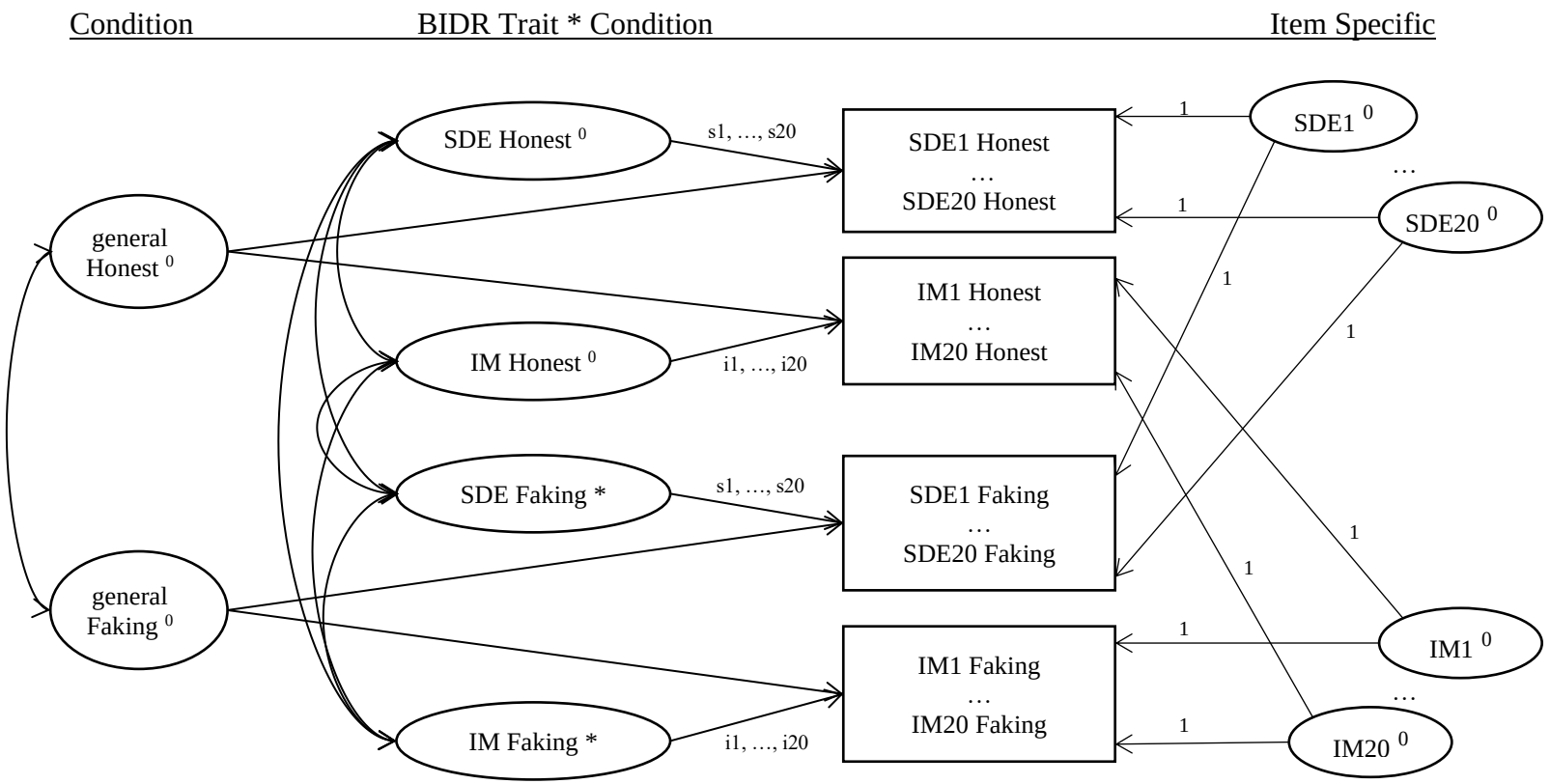
Table 7

Correlations between dichotomous observed scores, five-point observed scores, and trifactor mixture model factor scores

		Dichotomous				Five-point				Latent Mixture					
		hs	fs	hi	fi	hs	fs	hi	fi	hs	fs	hi	fi	h gen	f gen
Dichotomous	hs	1.00													
	fs	.36	1.00												
	hi	.21	.16	1.00											
	fi	.02	.49	.33	1.00										
Five-point	hs	.58	.15	.35	.09	1.00									
	fs	.13	.76	.20	.73	.17	1.00								
	hi	.13	.04	.85	.13	.35	.06	1.00							
	fi	.02	.57	.28	.92	.08	.76	.17	1.00						
Latent Mixture	hs	.53	.15	.38	.10	.94	.18	.40	.11	1.00					
	fs	.12	.74	.20	.81	.16	.97	.06	.84	.17	1.00				
	hi	.18	.01	.80	.05	.42	.01	.95	.10	.45	.02	1.00			
	fi	.07	.63	.27	.91	.11	.83	.15	.97	.14	.89	.09	1.00		
	h gen	.07	-.10	-.16	-.16	-.07	-.15	-.22	-.18	-.04	-.14	-.02	-.08	1.00	
	f gen	.12	-.10	-.06	-.22	.02	-.14	-.08	-.24	.04	-.15	.09	-.10	.79	1.00

Note. Dichotomous refers to the observed scores scored using the BIDR dichotomous scoring protocol; Five-point refers to the observed scores scored using the BIDR Likert scoring protocol with five scale points. fs=faking self-deceptive enhancement, fi=faking impression management, hs=honest self-deceptive enhancement, hi=honest impression management, h gen= general factor in honest condition, f gen= general factor in faking condition.

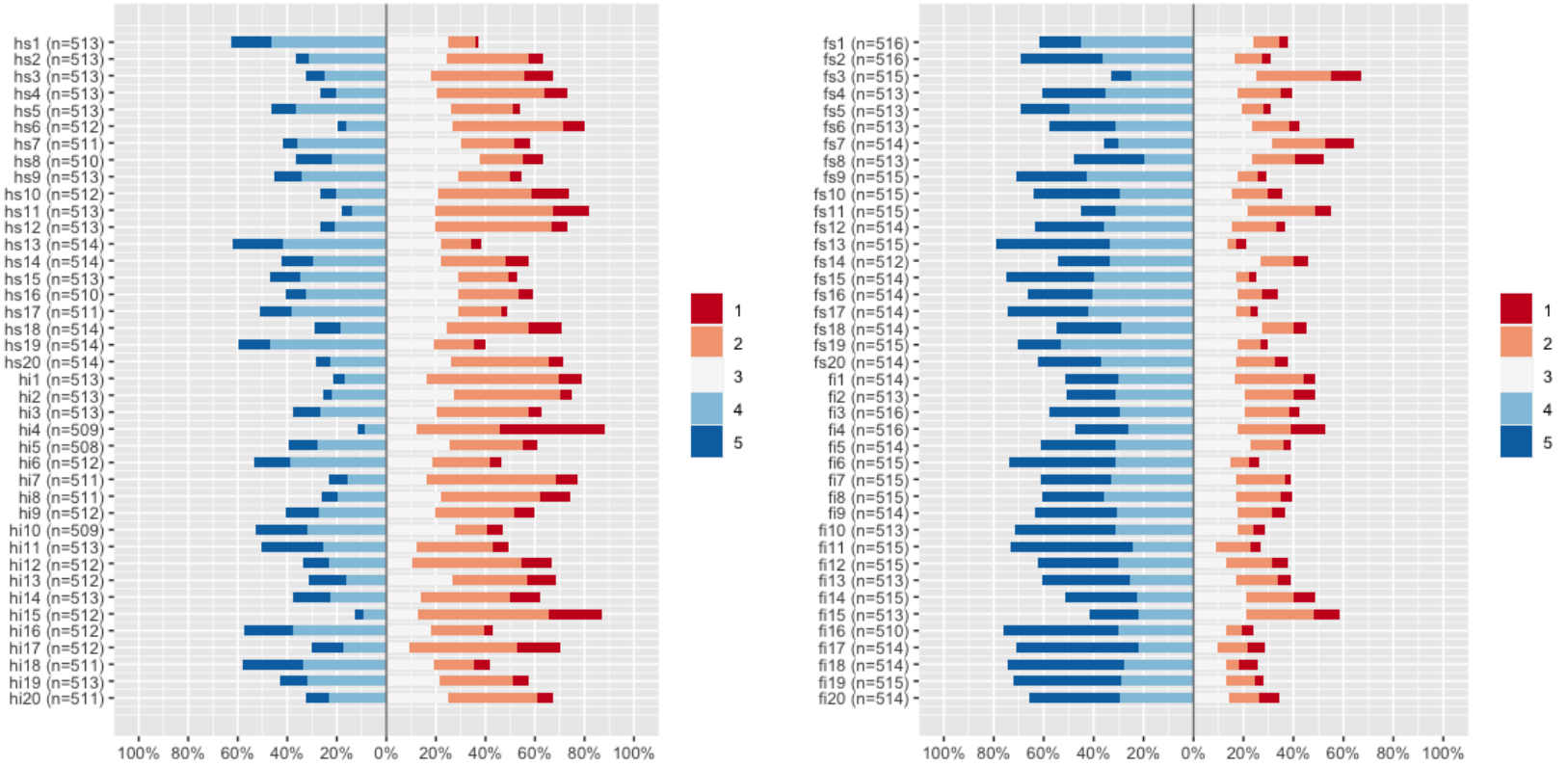
Figure 1. Diagram of the single class within subjects trifactor model



Note. Path coefficients are freely estimated unless they are constrained equal across conditions (indicated by labels i1…i20 and s1…s20) or fixed to 1 (indicated by label "1"). Small symbols above the latent variable names: 0 = mean fixed at zero; * = mean freely estimated.

Figure 2

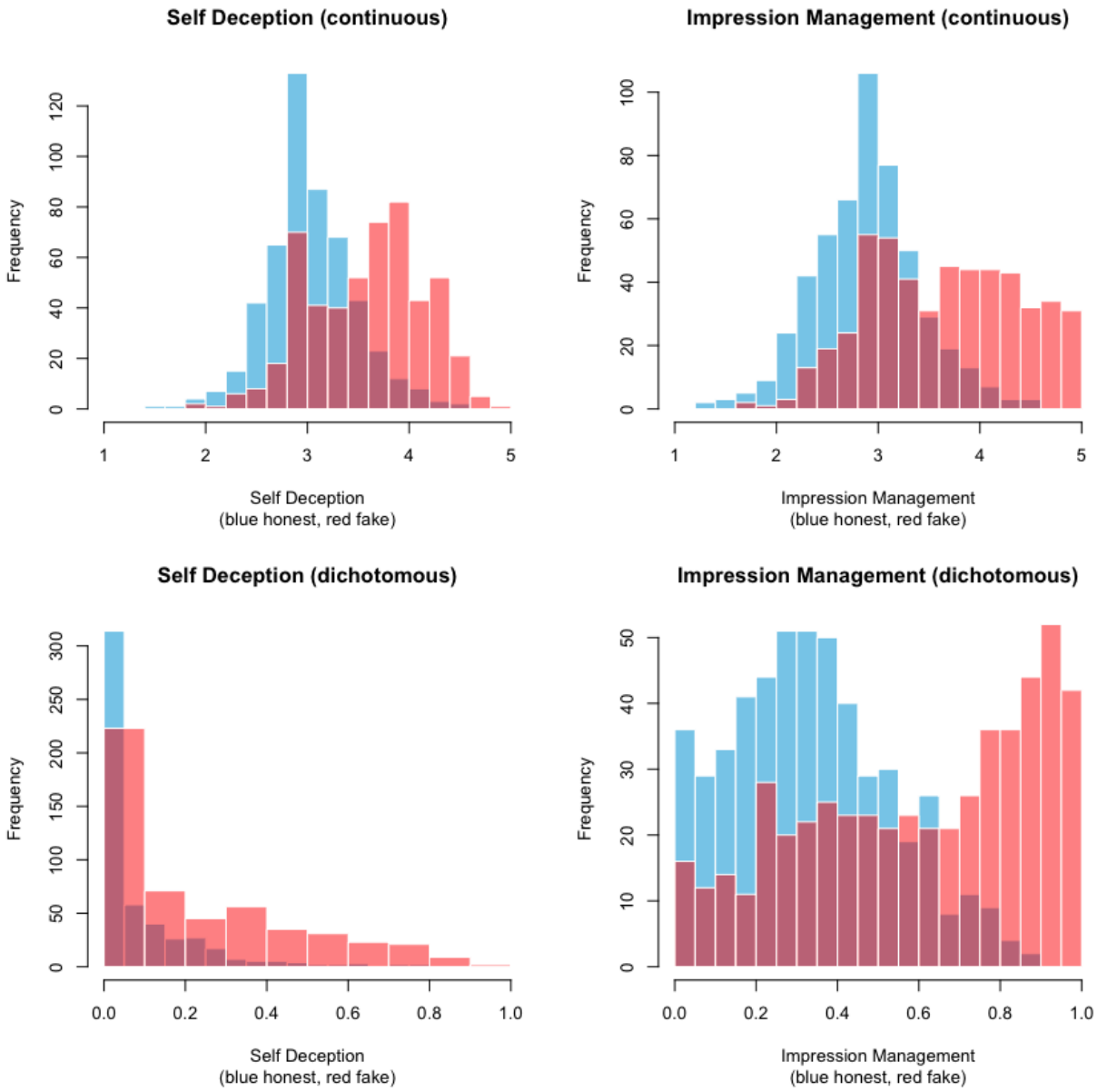
Observed item distributions (after reverse scoring) under honest (left panel) and faking (right panel) instructions (1-strongly disagree, 5=strongly agree).



Note. hs = honest condition for self-deceptive enhancement. hi = honest condition for impression management. fs = faking condition for self deceptive enhancement. fi = faking condition for impression management. The figure shows that the item distributions are much more heavily weighted towards strongly agreeing with socially desirable item content in the faking condition on the right compared to the honest condition on the left.

Figure 3

Observed score distributions for continuous and dichotomous scoring under honest and faking conditions



Note. In all cases the figure shows that in the faking condition score distributions shift higher (the red distributions that are shifted to the right in each panel). Observed variable continuous scoring using the full Likert scale range of 1 (strongly disagree) to 5 (strongly agree) resulting in the score distributions in the upper panels of this figure. These show the frequencies of participants with average scores across items within scales at values from 1 to 5 (i.e., using the original item scale). Observed variable dichotomous scoring scores each item according to the binary scoring protocol of the BIDR. This results in frequency distributions of scores ranging between 0 and 1 in the lower panels of this figure.