Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking

Eunike Wetzel[1,2], Susanne Frick[3], & Anna Brown[4]

[1]University of Vienna, Austria

[2]Otto-von-Guericke University Magdeburg, Germany

[3]University of Mannheim, Germany

[4]University of Kent, UK

Accepted version of article in press at *Psychological Assessment*

Date of acceptance: September 21, 2020

Author note:

This research was supported by a grant from the German Research Foundation (DFG) to Eunike Wetzel (WE 5586/2-1) as well as by the Elite Program for Postdocs of the Baden-Württemberg Stiftung and the Young Scholar Fund of the University of Konstanz. We thank Joschka Cremers, Theresa Falter, Celia Fürst, Veronika Held, Lars Hilbert, Clara Jupe, Mara Nuttelmann, Rachelle Sass, and Georg Schäfer for their help with data collection.

Correspondence concerning this article should be addressed to Eunike Wetzel, Department of Psychology, Otto-von-Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany. Email: eunike.wetzel@ovgu.de.

The following material is available on the Open Science Framework:

Pre-registration: https://osf.io/gk3js

Additional information on the questionnaire: https://osf.io/ft9ud/

Additional analyses: https://osf.io/7dmj9

Data: https://osf.io/q9uyp/

Does multidimensional forced-choice prevent faking? Comparing the susceptibility of the multidimensional forced-choice format and the rating scale format to faking

Abstract

A common concern with self-reports of personality traits in selection contexts is faking. The multidimensional forced-choice (MFC) format has been proposed as an alternative to rating scales (RS) that could prevent faking. The goal of this study was to compare the susceptibility of the MFC format and RS format to faking in a simulated high-stakes setting when using normative scoring for both formats. Participants were randomly assigned to three groups (total N = 1,867) and filled out the Big Five Triplets once under an honest instruction and once under a fake-good instruction. Latent mean differences between the honest and fake-good administrations indicated that the Big Five domains were faked in the expected direction. Faking effects for all traits were larger for RS compared to MFC. Faking effects were also larger for the MFC version with mixed triplets compared to the MFC version with triplets that were fully matched regarding their social desirability. The MFC format does not prevent faking completely, but it reduces faking substantially. Faking can be further reduced in the MFC format by matching the items presented in a block regarding their social desirability.

Keywords: forced-choice; rating scale; faking; social desirability; response format; Thurstonian item response model

Public significance statement: This study showed that it was harder for respondents to intentionally distort their responses when the questionnaire used a response format in which statements are ranked compared to a format in which statements are rated individually. The

forced-choice ranking format was especially effective at reducing faking when the statements presented together were about equally desirable.

Personality traits predict important life outcomes including occupational attainment, mortality, and divorce (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). For example, in the work domain, conscientiousness predicts job performance incrementally over cognitive ability (Schmidt & Hunter, 1998). Consequently, there is an interest in assessing personality traits in selection contexts, in particular to inform hiring decisions. In the clinical domain, self-reports of personality are used to inform the diagnosis and classification of personality disorders (Oltmanns & Widiger, 2018; Widiger & Samuel, 2005). However, a common concern is that self-reports of personality can be intentionally distorted by applicants and patients and that this in turn can bias selection and diagnostic decisions (Dilchert, Ones, Viswesvaran, & Deller, 2006; Goffin & Christiansen, 2003; Griffith, Chmielowski, & Yoshita, 2007; MacCann, Ziegler, & Roberts, 2011). For example, patients may distort their scores to achieve a particular diagnosis or compensation. With the most common response format used in self-report personality questionnaires - items presented as single stimuli that are responded to using a rating scale with several ordered categories - this concern is well-founded because it can be rather obvious to applicants which response category is the desirable choice. For example, with the conscientiousness item *I plan ahead* and the response options *strongly disagree – disagree – agree – strongly agree,* it is quite obvious to applicants that *agree* or *strongly agree* are the desirable responses. To mitigate this concern, the forced-choice format has been proposed as an alternative to rating scales that might be resistant or at least less susceptible to faking (e.g., Dilchert & Ones, 2011).

The forced-choice format

In the forced-choice (FC) format, two or more items are presented simultaneously to respondents (see example in Figure 1). In the simplest case, only pairs of items are presented and respondents select the item that describes them better. Other variants of the FC format present more than two items in an item block such as triplets or quads. Respondents then

either rank all items according to how well they describe them or select one that describes them most and one that describes them least. The items presented in one block either measure the same trait (unidimensional FC) or different traits (multidimensional FC; MFC). In our study, we applied a Big Five instrument with multidimensional triplets using a full ranking instruction, the Big Five Triplets (Wetzel & Frick, 2020).

When scored conventionally by using ranks as item scores, FC measures result in ipsative or partially ipsative scores that only allow intraindividual comparisons and distort correlation-based analyses such as reliability coefficients, factor analyses, and correlations with criteria (Brown & Maydeu-Olivares, 2018a; Hicks, 1970). The degree of ipsativity can be reduced by, for example, including negatively-keyed items, resulting in partially ipsative scores that nevertheless retain some ipsative constraints. However, recent model developments in the framework of item response theory allow deriving normative scores from FC data, thereby making interindividual comparisons possible. For an overview over the different item response models see Brown (2016). In this study, we applied the Thurstonian item response model (Brown & Maydeu-Olivares, 2011), an item response model that can readily be estimated using popular software[1]. In the Thurstonian item response model, ranks are coded into binary outcome variables from all pairwise comparisons between the items in the block. The model parameters can be estimated from the binary outcome variables using limited information methods (for details see Brown & Maydeu-Olivares, 2011; 2012; 2018a). The FC format has been shown to have similar or even better criterion-related validity than the rating scale (RS) format (Bartram, 2007; Lee, Lee, & Stark, 2018; Salgado & Táuriz, 2014; Wetzel & Frick, 2020; Wetzel, Roberts, Fraley, & Brown,

[1] For more information on how to estimate the Thurstonian item response model, see Brown and Maydeu-Olivares (2012). A tutorial and Excel macro for creating the Mplus syntax can be found on http://annabrown.name/software.

2016; Zhang et al., 2019), while generally having lower reliabilities given the same items as their single-stimulus counterparts (Brown & Maydeu-Olivares, 2018b).

The rationale behind using the FC format as a method of preventing faking is that when the items in a block are matched with respect to their desirability, applicants cannot fake all equally desirable and relevant items. In contrast, in the RS format, applicants only need to identify the most desirable response option for each individual item.

Previous research on faking in the forced-choice format

Previous research has found that the FC format is less susceptible to faking at the level of group mean differences (Christiansen, Burns, & Montgomery, 2005; Heggestad, Morrison, Reeve, & McCloy, 2006; Jackson, Wroblewski, & Ashton, 2000). For example, Christiansen et al. (2005) showed that means on conscientiousness and extraversion were elevated in both formats when participants were instructed to imagine applying for a sales position compared to when they filled out the questionnaire under an honest instruction. Importantly, however, the mean differences between honest and fake-good conditions were larger for the RS format (Cohen's *d* of 0.68 and 0.74) than for the MFC format (Cohen's *d* of 0.40 and 0.47). Furthermore, in a second study, Christiansen et al. found that MFC scores in the fake-good condition (applying for a customer service position) correlated moderately with performance ratings by supervisors whereas there was no relation for RS scores. Heggestad et al. (2006) also found smaller differences between honest and fake-good conditions for MFC compared to RS with the exception of conscientiousness, where the effect was similar for both formats. However, Heggestad et al. also conducted individual-level analyses, such as comparing the rank ordering of participants between a Big Five questionnaire completed in honest and faking conditions, and found that the correspondence in rank ordering was not better for MFC than RS.

A recent meta-analysis showed that the overall effect size for faking of FC measures was $d = 0.06$ (Cao & Drasgow, 2019), with effect sizes varying between 0 for neuroticism and openness and 0.23 for conscientiousness. These effect sizes are substantially smaller than those reported for RS measures, which ranged from 0.11 for extraversion to 0.45 for conscientiousness in a meta-analysis by Birkeland, Manson, Kisamore, Brannick, and Smith (2006). However, most of the 43 studies included in Cao and Drasgow's meta-analysis used pairwise comparisons and only one study used a full ranking format. Furthermore, most of the previous research used ipsative or partially ipsative scoring, rather than IRT-based normative scoring. Thus, more research on faking applying other variants of the FC format (e.g., triplets with full ranking) and normative scoring is needed.

The present study

The goal of this study is to compare the susceptibility of the MFC and RS format to explicit faking in a simulated high-stakes scenario when using normative scoring for both formats. Participants were randomly assigned to one of three versions of the Big Five Triplets (Wetzel & Frick, 2020): MFC-matched (all triplets consisted of items of equal desirability), MFC-mixed (7 triplets contained a desirable item in addition to two neutral or undesirable ones), and RS. They first filled out the respective Big Five Triplets (BFT) version under an honest instruction and later under a fake-good instruction. Our analyses address four research questions: 1) Does less faking occur in the MFC format compared with the RS format when both formats are scored normatively? 2) Does the degree of faking occurring in the MFC format depend on whether the items in a block are matched on desirability? 3) Does faking reduce criterion-related validity and – if yes – does the reduction in criterion-related validity differ across response formats? 4) Is the ability to fake successfully related to general intelligence?

This study extends previous research in several important ways: First, we consistently used normative scoring for both MFC and RS data whereas previous research confounded the response format with the scoring method - normative scoring was used for RS and ipsative or partially ipsative scoring for FC. Thus, it is unclear from previous research whether differences in the susceptibility to faking were due to the response format or the scoring method. As more and more MFC assessments apply IRT-based scoring, the effects of faking need to be investigated on the actual assessment scores. Thus, our study applies modern scaling methodologies to both MFC and RS to allow a fair comparison of their susceptibility to faking. Second, we designed two versions of our MFC instrument: one in which all triplets were carefully matched regarding their desirability and one in which some triplets contained items that differed in their desirability. This allowed us to test directly whether matching by desirability is a feasible strategy to reduce faking. In contrast, some previous studies mixed desirable and undesirable items within blocks and did not include a fully matched version. However, MFC questionnaires are argued to be less fakable only when items are matched. Third, we applied triplets with a full ranking instruction whereas most previous research applied forced-choice pairs. Triplets with full ranking are popular in MFC assessments because they yield more reliable scores than pairs when the number of items is held constant (Brown & Maydeu-Olivares, 2018a)[2]. Triplets might be harder to fake than pairs because with three statements, it might be harder to decide on the order of desirability than with pairs. Fourth, we also obtained data on criteria which allowed us to compare whether criterion-related validities for MFC and RS are differentially affected by faking. Importantly, most of

[2] Triplets with full ranking are more informative than pairs because the ranks can be broken down into three pairwise comparisons (Item A vs. Item B, Item A vs. Item C, Item B vs. Item C). This is how MFC data are analyzed in the Thurstonian item response model (Brown & Maydeu-Olivares, 2011), which appears to correspond to the underlying response process (Sass, Frick, Reips, & Wetzel, 2018). Thus, with triplets, three bits of binary information on participants' trait levels are obtained with each item block whereas only one bit of binary information is obtained with a pair when each item is presented only once.

these criteria were not measured with RS, thereby reducing common method bias between criteria and traits for RS. Fifth, a subsample of our participants filled out an intelligence test, allowing us to investigate whether the ability to fake successfully is related to intelligence in the MFC and RS format, a research question that has not been addressed with normatively scored MFC data.

Hypotheses

We preregistered the following hypotheses[3] on the Open Science Framework (https://osf.io/gk3js):

H1: Trait estimates on the Big Five with the faking instruction will differ from the ones with the neutral instruction in the direction of lower neuroticism, higher extraversion, higher agreeableness, higher conscientiousness, and higher openness for both the MFC and RS formats.

H2: The MFC format will be less susceptible to faking than the RS format; i.e., the differences predicted in H1 will be larger for the RS format.

H3: Within the MFC format, the differences predicted in H1 will be larger for the MFC-mixed socially desirable version than the MFC-matched socially desirable version.

Exploratory analyses

[3] We had originally planned to also investigate socially desirable responding in a low-stakes context with the following two preregistered hypotheses:
H1: In the MFC-mixed socially desirable (SD) version, the mean rank of the socially desirable option will be higher than the mean rank of the respective (neutral or socially undesirable) option assessing the same trait in the MFC-matched SD version.
H2: In the RS version, the mean rating of the socially desirable items will be higher than the mean rating of the respective items that are neutral or socially undesirable.
We later realized that social desirability could be confounded with item difficulty in these analyses and a higher endorsement of the socially desirable items could not be interpreted unambiguously as socially desirable responding, but could also reflect differences in item difficulty. We therefore decided not to report these analyses in the main text. However, they are available from osf.io/7dmj9 for interested readers. The hypotheses on faking were re-labeled H1 to H3.

In addition to testing the hypotheses described above, we also conducted two exploratory analyses. First, we compared the validity of the Big Five for predicting a number of criteria from different areas (e.g., social activities, health, and cognitive ability) between the honest and fake-good condition within each format. If scores on the Big Five are distorted by faking, predictive validity should be worse in the fake-good condition than in the honest condition. We then compared the difference in criterion-related validities between honest and fake-good across response formats. Second, we investigated whether the ability to fake successfully (faking ability) was related to general intelligence. Faking ability was operationalized 1) by the degree to which a respondent's ranks (MFC) or ratings (RS) aligned with the ideal ranks and ratings to fit the personality profile described in the faking scenario and 2) by the sum of each respondent's trait estimates in the fake-good condition.

Method

The study design and analysis plan were preregistered (https://osf.io/gk3js). This study was exempt from approval by an ethics committee.

Study design

Participants were randomly assigned to one of three response format groups: MFC-matched, MFC-mixed, and RS. The MFC-matched group filled out the original version of the Big Five Triplets (BFT; Wetzel & Frick, 2020), in which all triplets are matched with respect to their social desirability. The MFC-mixed group filled out an altered version of the BFT in which one item in the seven triplets containing items rated as socially undesirable or neutral was replaced by a socially desirable item (see below). The RS group filled out the items from BFT-matched and the additional socially desirable items from BFT-mixed presented as single stimuli with a rating scale. After being assigned to one of the three response format groups, participants first filled out the BFT in the respective format. Then, they responded to a number of questions relating to social activities, health, abilities, and other variables for a

study comparing validity between MFC and RS (Wetzel & Frick, 2020). The criterion variables will also be analyzed in this study to compare the predictive validity between the honest and the fake-good instruction. Next, participants filled out other personality questionnaires for the same study on validity. Then, they received a fake-good instruction and filled out the BFT again in the same version as at the beginning. After the end of the survey, participants in the laboratory subsample could optionally take a short intelligence test. Lastly, they were debriefed about the purposes of the data collection.

Sample

The data for this study came from two subsamples, one laboratory sample and one Internet access panel sample. The laboratory sample consisted of 1,042 persons from two German universities. They could choose between being remunerated with research participation credit or money (between 8 and 15 Euros depending on the length of the session). Participants who took the intelligence test could receive feedback on their scores. The Internet access panel sample (N = 1,217) was collected with Respondi, a German company. Respondi participants once register to the panel and are then invited to selected studies via email. For our study, only participants who were between 18 and 30 years old and whose first language was German were invited. Of the originally 1,217 participants, 91 were redirected without completing the study because they did not fulfill the age or language inclusion criteria or they participated after the quota for gender (50% female) was full. Participants in the access panel sample received 4 Euros for their participation. In both subsamples, we excluded some participants based on data quality checks (response time 2 *SD* below the average, incorrect responses to instructed response items), which led to the final sample sizes of 910 for the laboratory sample and 957 for the access panel sample, respectively. More detailed information on data exclusions and the demographic make-up of

the subsamples is available in Wetzel and Frick (2020)[4]. The complete sample across all three response format groups thus consisted of 1,867 persons. Of these, 593 participants were in the MFC-matched group, 652 were in the MFC-mixed group, and 622 were in the RS group. The demographic characteristics of participants in the three groups are provided in Table 1.

Measures

We only describe the measures relevant to this study in the following. For a description of other administered measures (personality questionnaires assessing the Big Five, HEXACO, and Dark Triad), see Wetzel and Frick (2020).

Big Five Triplets. The BFT was used to assess the Big Five domains neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. In the original version of the BFT (BFT-matched), all triplets are matched regarding their social desirability. Matching was based on social desirability ratings of all 213 items in the initial item pool obtained from a sample of 33 psychology students prior to test construction. Raters were told "Socially desirable means that the trait or behavior described in the statement fulfills societal norms and expectations." The instruction included three example items, one socially desirable, one neutral, and one socially undesirable item. For example, the item "In the bus I offer my seat to old people." was presented. It was then explained to participants that most people would consider this a behavior that fulfills societal norms and expectations and the item should consequently be rated as "socially desirable." The interrater reliability for the desirability ratings was excellent (ICC(1,k) = 0.99). For a more detailed description of the development of the BFT including a table of the items' mean desirability ratings, see Wetzel and Frick (2020) and https://osf.io/ft9ud/. The BFT in German and English can be

[4] Wetzel and Frick (2020) only analyzed data from the MFC-matched and RS groups from the first administration of the BFT with an honest instruction. There is no overlap in research questions or analyses with this study with the exception of the comparison of the predictive validity between honest and fake-good condition, which also uses the criterion-related validities from the honest instruction for MFC-matched and RS (see also below).

downloaded from https://osf.io/ft9ud/. In the BFT, neuroticism is assessed with 16 items, extraversion with 13 items, openness with 10 items, agreeableness with seven items, and conscientiousness with 14 items. BFT-matched showed good convergent validity with the Big Five Inventory (John, Donahue, & Kentle, 1991), except for agreeableness (see Wetzel and Frick, 2020, for a thorough investigation of construct and criterion-related validity). The matching procedure resulted in three types of triplets: triplets in which all items are socially undesirable, triplets with only neutral items, and triplets with only socially desirable items (see Figure 2 for examples of the three types). To construct BFT-mixed, we replaced one item in all the socially undesirable and neutral triplets with a socially desirable item. For example, in the triplet *I am not interested in other people's problems – I am a loner – I dislike myself*, we replaced the extraversion item (*I am a loner*) with *I warm up to others quickly*. In total, we replaced seven items: two on agreeableness, two on extraversion, two on conscientiousness, and one on neuroticism. Thus, BFT-matched and BFT-mixed are identical with the exception of these seven socially desirable items. Since items were presented as triplets, this implies that seven triplets differed between BFT-matched and BFT-mixed. For BFT-RS, the BFT-matched items plus the seven socially desirable items were presented individually (three per page) and participants were instructed to rate them on a four-point rating scale with the categories *strongly disagree*, *disagree*, *agree*, *strongly agree*. Empirical reliabilities for maximum a posteriori trait estimates on the Big Five from the honest instruction ranged from 0.68 (agreeableness) to 0.84 (conscientiousness) for BFT-matched, from 0.70 (agreeableness) to 0.84 (extraversion) for BFT-mixed and from 0.79 (agreeableness) to 0.91 (neuroticism and extraversion) for BFT-RS (see Table S1 in the supplemental online material).

Intelligence. Intelligence was assessed with the numeric module of a modular short intelligence test (Modularer Kurzintelligenztest, M-KIT; Dantlgraber, 2015), which is the

module that loads most strongly on the *g* factor. Scores on the numeric module ranged from 4 to 31 with a mean of 19.02 (*SD* = 4.68) and an omega total reliability of 0.79. Intelligence was operationalized as an observed sum score in all models.

Criteria. Table 2 shows the criteria and how they were assessed and Table S2 in the supplemental material shows descriptive statistics for the criteria. Participants responded to questions on criteria from five areas: social activities, health, charity, cognitive abilities, and other variables.

Fake-good instruction

For the fake-good instruction, we asked participants to imagine that they were applying to the Master's program of the Department of Psychology at a university in Germany. This scenario was realistic for the largest group in our laboratory sample who were psychology undergraduates (38%)[5]. The instruction read:

> *Universities have to select students into their programs based on certain criteria. Besides cognitive abilities, personality profiles also play a major role in the selection of candidates. Personality instruments such as the following are administered to candidates during the selection process. Please imagine that your goal is to be admitted as a student to the Master's program of the Department of Psychology at the University of Konstanz. The Department of Psychology is looking for students who are conscientious, reliable, extraverted, and emotionally stable. An ideal candidate would, for example, be punctual and assertive, and would complete assignments on time. He/she should be interested in people and be characterized by curiosity and creativity. Furthermore, the Department of Psychology is looking for students who*

[5] In Germany, almost all psychology undergraduate students have the goal of continuing to obtain a Master's degree because there is practically no job market for people with only a Bachelor's degree in psychology (Antoni, 2019).

are optimistic and purposefully pursue their academic goals. At the same time, they should be balanced, gregarious and helpful, and show compassion with others. Please fill out the following questionnaire in a way that fulfills these criteria in order to increase your chances of being admitted to the Master's program.

Thus, we instructed participants to fake all Big Five domains by making themselves more emotionally stable, more extraverted, more open, more agreeable, and more conscientious.

Analyses

Personality traits were modeled as latent variables in the framework of item response theory. We applied the Thurstonian item response model (Brown & Maydeu-Olivares, 2011; 2013) for MFC data and the graded response model (Samejima, 1969) for RS data. The models were estimated using unweighted least squares with mean- and variance-corrected Satorra-Bentler goodness-of-fit tests (ULSMV) in Mplus version 8 (Muthén & Muthén, 1998-2018). Observed scores were used to model criteria with the exception of life satisfaction, which was modeled as a latent variable.

Faking. To test our hypotheses on faking (H1 to H3), we first estimated latent mean differences between data from the honest instruction and data from the fake-good instruction for the three groups (MFC-matched, MFC-mixed, RS). This was done in two steps. First, we estimated item parameters in the appropriate model (Thurstonian item response model or graded response model) with the data from the honest instruction. In this first step, means of the latent traits were fixed to 0 and variances to 1. In the second step, we modeled the data from the fake-good instruction with item parameters fixed to those from the model with the data from the honest instruction. This was necessary to ensure that the measured constructs were on the same scale. Without fixed item parameters, it would not be possible to directly compare the latent traits from the first application of the BFT with an honest instruction to those from the second application with the fake-good instruction. This is because when

participants respond honestly, we should be measuring the Big Five. However, when they fake, the latent traits reflect what participants consider to be the ideal responses in the context of the faking scenario. Thus, if the goal is to compare mean levels on the latent traits between the two applications, it is essential to ensure that the measured traits are on the same scale, which is achieved by fixing the item parameters in the faking model to those from the honest model[6]. Further, this procedure mimics applied contexts, where traits for individuals are estimated from item parameters obtained a priori. In the model with the data from the fake-good instruction, means and variances of the latent traits were freely estimated and the means directly correspond to the difference between the honest and the fake-good instruction.

To test H1 on whether the Big Five were faked in the expected direction, we tested whether the size of the latent mean difference for each trait was at least small according to Cohen's (1988) criteria ($d \geq |0.20|$). H2 and H3 were tested by comparing the latent mean differences in *d* between MFC-matched and RS (H2) and MFC-matched and MFC-mixed (H3), again applying Cohen's criteria for at least small differences.

Comparison of predictive validity. As an additional (not preregistered) exploratory analysis, we compared the predictive validity of the Big Five between honest and fake-good instructions within each response format and descriptively also across response formats. First, five experts in the area of personality psychology rated for which of the 20 criteria listed in Table 2 they expected a latent correlation > |.15|, with .15 as the cut-off for a small latent correlation (Gignac & Szodorai, 2016). We used the modal ratings as hypothesized criterion-related correlations. Criterion-related correlations were estimated in separate models for the

[6] Note that this goal would not be achieved by imposing measurement invariance across the data from the two instructions because the measured constructs differ fundamentally. It is therefore unlikely that even configural invariance would exist and constraining item parameters to equality across the two instructions would result in item parameter estimates that cannot be interpreted. Investigating measurement invariance could be used as a tool to investigate faking at the item-level, but would probably distort comparability between honest and fake-good person scores.

data from the honest instruction and the data from the fake-good instruction. In these models, we also included the other (non-hypothesized) correlations between the Big Five and the criteria in order to compare them with the hypothesized criterion-related correlations. In this analysis, the item parameters in the faking models were also fixed to those from the honest models. We then computed the average criterion-related validity and the average across the other correlations for each instruction × response format combination. We only compare the correlations descriptively because we did not formulate any a priori hypotheses for this research question.

Faking ability. Lastly, we were interested in whether participants' ability to fake successfully (their *faking ability*) was related to intelligence. We operationalized faking ability in two ways. First, as the Mahalanobis distance between participants' fake-good Big Five profile and an ideal Big Five profile based on expert ratings and second, as the sum of the maximum a posteriori (MAP) estimates on the Big Five from the fake-good administration. With respect to the first method, to obtain the ideal profile, experts were instructed to rank/rate the items in the way that best fulfilled the faking instruction. They were allowed to consult the faking instruction the entire time while providing rankings/ratings for the ideal profile. We then used the modal rank/rating across experts to obtain the final ideal profile. In the MFC-format, this sometimes resulted in multiple modes for an item (e.g., if three raters ranked it second and three raters ranked it third)[7]. If the ranks for the other two items in the triplet were unambiguous, the third item received the remaining rank. For two triplets, there was no clear mode for more than one item. The authors resolved these cases by discussion. MAP estimates for the Big Five were obtained from a model with

[7] Three experts rated/ranked the ideal profile for RS and the undesirable and neutral triplets in MFC-matched, but six experts ranked the ideal profile for MFC-mixed (including the overlapping triplets with MFC-matched). Thus, multiple modes could only occur in the MFC-format. Multiple modes occurred for 14 items.

the fake-good data in which item parameters were fixed to those from the respective honest-instruction model. To obtain MAP estimates for the expert-specified ideal profile, we used the item and trait parameters from the fake-good model estimated with only the real participants. We then calculated the Mahalanobis distance between the participants' MAPs and the ideal case MAPs and correlated it with the participants' scores in the intelligence test. The second way in which we operationalized faking ability was as the sum of the MAP estimates on the Big Five from the fake-good administration, with neuroticism reversed. This sum was then correlated with intelligence test scores.

Besides Mplus, we also used the following R (R Core Team, 2018) packages in data analysis: BSDA (Arnholt & Evans, 2017), compute.es (Re, 2013), lattice (Sarkar, 2008), MplusAutomation (Hallquist & Wiley, 2018), and psych (Revelle, 2018).

Results

Faking

We only evaluated the fit of the models with the data from the honest instruction because the models with the data from the faking instruction had fixed item parameters. According to the RMSEA, the MFC-matched and MFC-mixed models showed an excellent fit (0.036 and 0.038, respectively) and the RS model showed an acceptable fit (0.062). The SRMR and CFI indicated an acceptable to slightly below acceptable fit for all models (see Table S3). First, we checked whether the Big Five were faked in the expected direction. This was the case for all Big Five domains and all response format groups except agreeableness in MFC-matched (see Figure 3 and Table 3), overall confirming H1. Thus, when instructed to present themselves favorably in order to be admitted to the Master's program in psychology, participants distorted their responses to appear more emotionally stable, more extraverted, more open, more agreeable, and more conscientious compared to when they filled out the BFT honestly. Effect sizes for the hypothesized changes ranged from $d = 0.24$ (95% CI =

[0.12; 0.35]) for openness to experience in MFC-matched to $d = –1.38$ (95% CI = [–1.51; –1.26] for neuroticism in RS[8]. The variance in trait estimates was smaller under the fake-good instruction than the honest instruction for MFC for all traits except conscientiousness in MFC-matched. For RS, the opposite was the case with larger variances under faking than honest responding for all traits except agreeableness. We additionally checked how much variance there was in the ranking patterns for MFC and in the response distributions for RS under the honest versus fake-good instruction. The variance was smaller under the fake-good instruction for the majority of triplets (17 out of 20 for MFC-matched and 18 for MFC-mixed) and items (54 out of 67 for RS). Furthermore, correlations between MAPs from the honest instruction and the fake-good instruction were only low to moderate (range from .13 to .36 for MFC-matched, .14 to .23 for MFC-mixed, and .19 to .31 for RS). Thus, when faking, participants showed less variance in their responses and changed their rank ordering.

The latent mean differences between honest and fake-good were larger for RS than MFC-matched for all Big Five traits with moderate to large effect sizes (absolute differences in Cohen's *d* ranged from 0.49 for extraversion to 0.93 for agreeableness), which confirms H2. Next, we compared MFC-matched and MFC-mixed. For extraversion, agreeableness, and conscientiousness, the traits for which socially undesirable or neutral items were replaced with socially desirable ones, the latent mean differences between honest and fake-good were larger for MFC-mixed compared to MFC-matched with small effect sizes for extraversion (0.38) and conscientiousness (0.21) and a moderate effect size for agreeableness (0.67). For

[8] Because the scenario was especially relevant for psychology students and Ziegler (2007) showed that psychology majors tend to fake other aspects than students from other majors, we also did an exploratory investigation of the faking effects in the subsample of psychology students. Due to the small sample sizes (*N*s between 107 and 120 for the three groups), we did not estimate new models, but rather used MAPs from the models with the full sample. Consistently across the three formats and in line with Ziegler's results, psychology students faked neuroticism and conscientiousness more strongly than the full sample (e.g., $d = –1.38$ for neuroticism and $d = 0.92$ for conscientiousness for MFC-matched). There were no systematic differences for the other traits (see Table S4).

neuroticism and openness, which were identical between MFC-matched and MFC-mixed, latent mean differences between honest and fake-good administrations went in the direction of stronger faking for MFC-mixed than MFC-matched, but were negligible in terms of effect sizes. Thus, overall, H3 was confirmed.

In an exploratory analysis suggested by a reviewer, we additionally compared the actual faking levels (Cohen's *d* between honest and fake-good instruction) with ideal faking (Cohen's *d* between honest instruction and the expert-rated ideal profile) using MAPs. Ideal faking thus represents the amount of change in scores we would expect if participants did not respond in accordance with their own trait levels at all, but only followed the faking instruction. Absolute Cohen's *d* values were larger for ideal faking than actual faking in almost all cases (see Table S5). For example, for conscientiousness, Cohen's *d* for actual faking was 0.47 for MFC-matched, 0.63 for MFC-mixed, and 1.62 for RS, whereas Cohen's *d* for ideal faking was 0.74 for MFC-matched, 0.78 for MFC-mixed, and 3.06 for RS. In addition, Cohen's *d* values for ideal faking were substantially smaller for MFC than RS, indicating that the MFC format was also less susceptible to faking when experts filled out the questionnaire following the instruction meticulously.

Criterion-related validity

Latent criterion-related correlations ranged from 0 for neuroticism with smoke (yes/no) in MFC-matched (honest instruction) to –.67 for neuroticism with life satisfaction in MFC-mixed (honest instruction; see Table 4). Number of cigarettes a day was removed due to estimation problems which can probably be attributed to the small sample sizes for this criterion (between 72 and 74 for the three groups). The full correlation table is depicted in supplementary Table S6. The average criterion-related correlation across all expert-hypothesized relations under the honest instruction was .19 (95% CI = [.12; .25]) for MFC-matched, .26 [.18; .33] for MFC-mixed, and .22 [.14; .29] for RS. Under the fake-good

instruction, average criterion-related correlations were reduced to around .10 for all groups[9]. Average criterion-related correlations under the honest instruction were larger than the average over the other (non-hypothesized) correlations for all groups and the average of the other correlations did not differ between the honest and the fake-good instruction (all averages were between .09 and .12). In sum, instructed faking reduced the criterion-related validities for all response format groups about equally.

Faking ability and intelligence

The Mahalanobis distance between participants' fake-good Big Five profile and the ideal Big Five profile correlated at approximately $r = 0$ with their score in the short intelligence test for MFC-matched ($r = –0.01$) and MFC-mixed ($r = 0.03$). In the RS group, the correlation was slightly below small (cutoff = |.10| for observed correlations; Gignac & Szodorai, 2016) at $r = –0.09$. The sum of the MAP estimates from the fake-good administration did not correlate with intelligence test scores in the MFC-matched group ($r = –0.04$), but showed a small correlation in the MFC-mixed group ($r = 0.11$) and an almost small correlation in the RS group ($r = 0.09$). Thus, the results on whether faking ability is related to intelligence were inconclusive.

Discussion

In this study, we compared the susceptibility of the MFC and RS format to faking in a simulated high-stakes setting. In the following, we will first discuss our results on faking. Then, we will discuss the important issue of matching items regarding their desirability. Lastly, we will discuss the effects of faking on criterion validity in the two formats before noting limitations of our study and future directions.

[9] We checked whether results differed when the item parameters were freely estimated in the models with the faking data. This was not the case. Average criterion-related validities were practically identical as those reported above at .10 for MFC-matched, .08 for MFC-mixed, and .09 for RS.

Intentional faking in a simulated high-stakes setting

Our hypotheses regarding the latent mean differences between honest and fake-good conditions were confirmed. Thus, participants adhered to the fake-good instruction and described themselves as more emotionally stable, more extraverted, more open, more agreeable, and more conscientious compared to when they were responding honestly. The only exception to this pattern was that there was no mean difference between honest and fake-good for agreeableness in MFC-matched. This result is possibly due to issues with the construct validity of agreeableness in the matched version of the BFT (see Wetzel & Frick, 2020). The effect sizes of our latent mean differences between honest and fake-good were overall moderate to large, which is typical for simulated faking designs whereas effect sizes would be expected to be smaller in real settings (Holden & Book, 2011; Smith & Ellingson, 2002).

One difference to previous research on faking of the Big Five is that in our study neuroticism was faked more strongly than conscientiousness whereas in meta-analyses on faking in the MFC format and the RS format, conscientiousness was the trait with the strongest faking effects (Birkeland et al., 2006; Cao & Drasgow, 2019). The faking instruction included three keywords related to neuroticism and five related to conscientiousness, indicating that an imbalance in the faking instruction was probably not the reason. One possible reason is that there were more neuroticism items than conscientiousness items (16 vs. 14), giving participants more opportunity to fake neuroticism.

The comparison of the two MFC versions with the RS version showed that faking effects were larger in the RS version for all Big Five domains. Thus, in line with previous research (Cao & Drasgow, 2019; Christiansen et al., 2005; Heggestad et al., 2006; Jackson et al., 2000), we found that the MFC format, while not being completely faking-resistant, is substantially less susceptible to faking than the RS format. Therefore, in assessment contexts

in which faking is a concern, such as personnel selection or clinical diagnosis, using the MFC format can be recommended. This statement can be further qualified by our finding from the comparison between the MFC-matched and the MFC-mixed version of the BFT: The BFT version that contained only fully matched triplets regarding the items' desirability (MFC-matched) showed smaller latent mean differences between honest and fake-good instructions than the BFT version that contained seven mixed triplets (MFC-mixed). Thus, careful matching of the items presented in triplets regarding their social desirability can further reduce faking in the MFC format (see also Cao & Drasgow, 2019). We now turn to the issue of how this matching can be achieved.

Matching of items with respect to desirability

Different methods exist for matching items regarding their desirability. One method is to match items based on their item means from an administration in the RS format (Jackson et al., 2000; Watrin, Geiger, Spengler, & Wilhelm, 2019). In this case, the item means are interpreted as the popularity and therefore desirability of the items. Another method is to obtain ratings of the desirability of the items and to match items with similar desirability (Christiansen et al., 2005; Converse et al., 2010; Edwards, 1953; Heggestad et al., 2006; Jackson et al., 2000). In our study, we chose the latter approach and obtained explicit ratings of the social desirability of all items in our pool prior to constructing the BFT. Then, we combined items to triplets that had the same social desirability rating. While matching based on item means might be simpler, it neglects that an item's mean does not purely reflect its desirability, but rather prevalence of the behavior, which is influenced by many factors. Other research has shown that applicants are able to identify selection criteria (Klehe et al., 2012) and two items with equivalent means ('difficulties' in item response theory terms) but from different traits can differ in desirability when one is more relevant to the assessment.

One disadvantage of both matching methods is that they assume that item desirability stays the same when items are presented together in one block. However, items having similar desirability in the RS format or identical social desirability ratings does not guarantee that they will be perceived as equally desirable when they are presented together in one block. In the context of our method, it is possible that items that were all individually rated as socially desirable, such as those in the right triplet in Figure 2, are perceived as differing in their social desirability when presented simultaneously. As Feldman and Corah (1960, p. 480) put it: "single statements may acquire contextual meaning when paired; hence their SD [social desirability] values may be somewhat altered." The process of responding to MFC triplets involves weighing the items in the triplet against each other before assigning them ranks (Sass et al., 2018), and fine-grained differentiations regarding desirability could be a part of this comparison (Kahneman, 2011). As Lin and Brown (2017) showed, differences in perceived social desirability can occur with different arrangements of items to blocks and this can influence the items' psychometric properties. Thus, the question of how to best match items in MFC test construction is still unanswered. Other methods could be explored such as obtaining comparative ratings of the items' social desirability prior to assembling the triplets (e.g., presenting different combinations of pairs of items and asking experts to rate whether one is more desirable or whether they are equally desirable). Future research could therefore compare different methods of matching items by desirability and investigate which method achieves the best match, for example in terms of reducing faking effects.

It has been suggested that matching by desirability in the MFC format is only possible with equally-keyed items because with mixed-keyed items, participants will easily realize which items are positively keyed and prefer those (Bürkner, Schulte, & Holling, 2019). To our knowledge, no empirical evidence has been offered for this claim. Following test construction guidelines for the MFC format (Brown & Maydeu-Olivares, 2011), we

combined positively and negatively-keyed items in all BFT triplets but one. This was possible because positively and negatively-keyed items were rated as having equal desirability. Our empirical data indicate that participants did not always prefer the positively-keyed items. For example, in the triplet consisting of *I tend to be very particular about things – I stay in the background – I have a vivid imagination*, the negatively-keyed extraversion item (I stay in the background) received a median rank of 2 under the honest instruction and was ranked first by 33% of the participants. This indicates that it is possible to match by desirability even with mixed-keyed items. Nevertheless, future research could further investigate matching by desirability with mixed-keyed items and whether ranking patterns are related to the keying of the items in a block.

Another difficulty with item matching is that socially desirable responding and faking are context-dependent. The faking profile will differ depending on the specific faking instruction. For example, Pauls and Crost (2005) found markedly different faking profiles depending on whether participants were instructed to generally fake good, imagine they were applying for a position as a manager, or imagine they were applying as a nurse. Thus, the more far-reaching question is whether it is at all possible to achieve a matching that is valid across contexts, or whether it would be necessary to adapt the matching of items to each specific context. Converse et al. (2010) showed that different instructions in obtaining desirability ratings (general desirability, general job applicant, real estate job applicant, police officer job applicant) resulted in different instrument versions. However, when the faking instruction only matched for the police officer job applicant scenario, instrument versions based on desirability ratings for the other two job applicant scenarios did not perform notably worse than the police officer instrument version (Converse et al., 2010). In contrast, stronger faking effects were found for the instrument version based on general desirability ratings. Thus, when the goal is to apply an instrument in selection contexts, raters should evaluate the

desirability of the items in a selection context also, though it does not appear to need to be job-specific. In clinical settings, faking bad is more of a concern than faking good. First evidence indicates that the psychological processes underlying faking good and faking bad may differ (Bensch, Horstmann, Greiff, & Ziegler, 2019). More research is needed to illuminate the psychological processes underlying faking good and bad and whether general desirability matching is able to reduce faking bad. In our study, we applied an instrument that was not specifically designed for use in selection contexts, but rather more generally for personality assessment (Wetzel & Frick, 2020). The instruction for obtaining the social desirability ratings was rather broad and defined desirability as fulfilling societal norms and expectations. In contrast, the faking instruction painted the scenario of applying to a Master's program in psychology and contained very specific trait descriptions that could be classified as eliciting agency management ("purposefully pursue their academic goals") and communion management ("show compassion with others") in terms of Paulhus' (2002) two-tier system of socially desirable responding. Nevertheless, faking was substantially reduced in the MFC format. It is possible that the MFC version would have been even more resistant to faking if the desirability ratings had been obtained for a selection context. Thus, when constructing an MFC instrument, the situations in which it will be applied should inform item matching.

Predictive validity

It has been argued that faking does not influence criterion validity (Ones & Viswesvaran, 1998; Ones, Viswesvaran, & Reiss, 1996), though other studies have found a detrimental effect of faking on criterion validity (Douglas, McDaniel, & Snell, 1996; Holden, 2007; Holden, Wood, & Tomashewski, 2001; Ziegler & Bühner, 2009). One reason for the inconsistent findings in previous research might be the setting and the base rate of faking associated with it (Holden & Book, 2011). In natural settings, only part of the respondents

will fake whereas in induced faking settings, all respondents compliant with the instruction can be expected to fake. In our instructed faking design, criterion-related validities were consistently worse under the faking instruction for both the MFC and the RS format. Furthermore, for the faking data, criterion-related correlations were indistinguishable from the other, non-hypothesized correlations between traits and criteria. Average criterion-related validities were very similar in all three groups (MFC-matched, MFC-mixed, RS), and the degree to which criterion-related validities were reduced in the fake-good condition was also approximately equal. Thus, the effects of faking on criterion validity were similar for the MFC and the RS format and the MFC format therefore did not show an advantage compared to RS, despite the overall lower degree of faking. Our instrument was the same length in both formats, thereby necessarily yielding lower reliability for the MFC format (Brown & Maydeu-Olivares, 2018b), though this should not have affected *latent* correlations with criteria. Nevertheless, future research could compare the effects of faking on criterion-related validity with MFC and RS instruments that have similarly high reliability. Especially for clinical settings, investigating predictive validity is important because personality scores are not used in isolation (such as for rank-ordering persons), but rather for predicting a diverse range of behaviors and outcomes.

Faking ability

A previous study by Vasilopoulos, Cucina, Dyomina, Morewitz, and Reilly (2006) found a relationship between cognitive ability and personality scores from a forced-choice measure in the faking condition, but not in the honest condition. In our study, faking ability was uncorrelated with intelligence when we operationalized it as the Mahalanobis distance between the participants' fake-good profile and the expert-rated ideal profile. When we used the sum of the MAP estimates (with neuroticism reversed) from the fake-good administration instead, we found no correlation for MFC-matched, but a small correlation for MFC-mixed

and RS. It could be argued that in selection contexts in which applicants are selected in the order of their scores, the sum of the MAP estimates is a better measure of faking ability because it reflects whether participants adhered to the principle that the higher the score, the better. In the two versions that could be faked effectively, RS and MFC-mixed, the small correlation indicates that more intelligent participants tended to obtain higher scores than less intelligent participants and would therefore have a higher probability of being selected. The null correlation in MFC-matched is further evidence that the matching by desirability was successful. However, as these were non-preregistered, exploratory analyses, it would be important to replicate them before drawing any conclusions. In addition, it is possible that there was some variance restriction in intelligence scores because the largest group in our laboratory sample were psychology students, who are selected into the program by high school GPA and the cut-off is quite high due to a limited number of places. Furthermore, the intelligence test was an optional part of the laboratory session, making it possible that some self-selection took place. Thus, more research investigating the association between faking ability and intelligence in more heterogeneous samples is needed.

Limitations and future directions

One important limitation of our study is that participants were not really applying to the Master's program in psychology, but that it was a simulated high-stakes situation. However, the scenario was chosen to be realistic for psychology undergraduates who formed 38% of our laboratory sample or 19% of our full sample. Previous research, on the other hand, often used lower prestige jobs that do not require a college education (e.g., assembler and customer service representative in Study 3 in Christiansen et al., 2005) with undergraduate samples, leading to a lack of fit between the sample's characteristics and the job descriptions in the faking scenario. Our goal was to avoid this. Nevertheless, future research on real applicants or patients is needed. It is very likely that smaller faking effects

for the MFC format will be found in real settings as has been the case for rating scales (Holden & Book, 2011; Smith & Ellingson, 2002).

In our study, participants were instructed to fake all traits. Past research has shown that participants tend to fake only the traits that they consider relevant, for example, for a particular job application (Birkeland et al., 2006; Cao & Drasgow, 2019; Furnham, 1990; Pauls & Crost, 2005). If some traits are classified as irrelevant by applicants, it is possible that they will be better able to focus on faking the relevant traits, thereby potentially reducing the advantage of the MFC format compared to the RS format. Thus, future research could induce faking on only some traits in order to investigate the susceptibility of the MFC format to faking. In addition, future research could vary the number of traits and investigate whether faking effects are reduced when the number of traits is larger – perhaps combined with conditions in which varying numbers of traits are relevant to the assessment context. Another factor that can influence faking effects is interindividual differences in the tendency to fake. In our study, these were presumably not a relevant issue because we used an induced faking design and the faking instruction detailed very clearly which traits to fake in which direction. However, in natural settings, applicants' tendency to fake can play an important role. Pavlov, Maydeu-Olivares, and Fairchild (2019) showed using a regression-based moderation analysis that faking was only reduced for the forced-choice format at high faking tendency levels, though they did not match the forced-choice items regarding desirability. This implies that future research conducted in natural settings should also take participants' propensity for faking into account. Our faking instruction was very detailed and specific in order to reduce variance in participants' interpretation or influence of prior knowledge on which traits should be faked in which direction. In other contexts, test takers may differ in their knowledge about a job or the criteria of a diagnosis. In these situations, it would also be interesting to investigate whether faking ability is related to crystallized intelligence.

The degree of faking occurring in the MFC format may also depend on the number of items presented in a block (pairs, triplets, tetrads, pentads) and the instruction (full ranking, most like me/least like me). Our study extended previous research by employing triplets with a full ranking instruction. Triplets have the advantage that they achieve a good balance between information and cognitive load. Most previous research on faking in the MFC format either used pairs and instructed participants to choose the response option that described them best (Christiansen et al., 2005; Converse et al., 2010) or used quads and instructed participants to choose the statement that was most like them and the one that was least like them (Heggestad et al., 2006; Jackson et al., 2000). With more statements presented simultaneously and a full ranking instruction, it might be harder for respondents intending to fake to figure out the order of desirability. However, at the same time, it might become more cognitively taxing for honest participants to describe themselves accurately. Thus, future research could investigate whether other MFC format variants, such as quads with a full ranking task, are able to reduce faking effectively while still being manageable for honest respondents. Our study used a within-subjects design in which participants filled out the questionnaire both under an honest instruction and under a fake-good instruction. It is possible that some participants felt the need to change their responses across conditions, though our results are consistent with other studies that employed a between-subjects design (e.g., Christiansen et al., 2005; Heggestad et al., 2006).

Conclusion

In conclusion, we found that the MFC format was less susceptible to faking than the RS format when both formats were scored normatively. This was especially the case when items within triplets were matched regarding their social desirability.

References

Antoni, C. H. (2019). Zur Lage der Psychologie [On the state of psychology]. *Psychologische Rundschau, 70*(1), 4-26.

Arnholt, A. T., & Evans, B. (2017). BSDA: Basic Statistics and Data Analysis. (Version 1.2.0.). Retrieved from https://CRAN.R-project.org/package=BSDA

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*(3), 263-272.

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

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. doi:10.1111/j.1468-2389.2006.00354.x

Brown, A. (2016). Item response models for forced-choice questionnaires: A common framework. *Psychometrika, 81*(1), 135-160. doi:10.1007/s11336-014-9434-9

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

Brown, A., & Maydeu-Olivares, A. (2012). Fitting a Thurstonian IRT model to forced-choice data using Mplus. *Behavior Research Methods, 44*(4), 1135-1147. doi:10.3758/s13428-012-0217-x

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

Brown, A., & Maydeu-Olivares, A. (2018a). Modeling of forced-choice response formats. In P. Irwing, T. Booth, & D. Hughes (Eds.), *The Wiley Handbook of Psychometric Testing*. London, UK: John Wiley & Sons.

Brown, A., & Maydeu-Olivares, A. (2018b). Ordinal factor analysis of graded-preference questionnaire data. *Structural Equation Modeling, 25*(4), 516-529. doi:10.1080/10705511.2017.1392247

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

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

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*(3), 267-307. doi:10.1207/s15327043hup1803_4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum.

Converse, P. D., Pathak, J., Quist, J., Merbedone, M., Gotlib, T., & Kostic, E. (2010). Statement desirability ratings in forced-choice personality measure development: Implications for reducing score inflation and providing trait-level information. *Human Performance, 23*(4), 323-342. doi:10.1080/08959285.2010.501047

Dantlgraber, M. (2015). *M-KIT: Modularer Kurzintelligenztest* [M-KIT: Modular Short intelligence test]. Bern, Switzerland: Hogrefe.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction with Life Scale. *Journal of Personality Assessment, 49*(1), 71-75. doi:10.1207/s15327752jpa4901_13

Dilchert, S., & Ones, D. S. (2011). Application of preventive strategies. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 177-200). New York: Oxford University Press.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: Born to deceive, yet capable of providing valid self-assessments? *Psychology Science, 48*(3), 209-225.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996). The validity of non-cognitive measures decays when applicants fake. In J. B. Keyes & L. N. Dosier (Eds.), *Proceedings of the Academy of Management* (pp. 127-131). Madison, WI: Omnipress.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *The Journal of Applied Psychology, 37*, 90-93.

Feldman, M. J., & Corah, N. L. (1960). Social desirability and the forced choice method. *Journal of Consulting Psychology, 24*, 480-482.

Furnham, A. (1990). Faking personality questionnaires - Fabricating different profiles for different purposes. *Current Psychology: Research & Reviews, 9*(1), 46-55. doi:10.1007/Bf02686767

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. doi:10.1016/j.paid.2016.06.069

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

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*(3), 341-357.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling, 25*(4), 621-638. doi:10.1080/10705511.2017.1402334

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*(1), 9-24.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167-&. doi:10.1037/h0029780

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Canadian Journal of Behavioural Science, 39*(3), 184-201. doi:10.1037/cjbs2007015

Holden, R. R., & Book, A. S. (2011). Faking does distort self-report personality assessment. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 71-86). New York: Oxford University Press.

Holden, R. R., Wood, L. L., & Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity? *Journal of Personality and Social Psychology, 81*(1), 160-169.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*(4), 371-388. doi:10.1207/S15327043hup1304_3

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory-Versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Kahneman, D. (2011). *Thinking, fast and slow*. London, UK: Allen Lane.

Klehe, U. C., Kleinmann, M., Hartstein, T., Melchers, K. G., Konig, 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

Lee, P., Lee, S., & Stark, S. (2018). Examining validity evidence for multidimensional forced choice measures with different scoring approaches. *Personality and Individual Differences, 123*, 229-235. doi:10.1016/j.paid.2017.11.031

Lin, Y., & Brown, A. (2017). Influence of context on item parameters in forced-choice personality assessments. *Educational and Psychological Measurement, 77*(3), 389-414.

MacCann, C., Ziegler, M., & Roberts, R. D. (2011). Faking in personality assessment. In M. Ziegler, C. MacCann, & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 309-329). New York: Oxford University Press.

Muthén, L. K., & Muthén, B. O. (1998-2018). Mplus [Computer software]. Los Angeles, CA: Muthén & Muthén. Retrieved from www.statmodel.com

Oltmanns, J. R., & Widiger, T. A. (2018). A self-report measure for the ICD-11 dimensional trait model proposal: The Personality Inventory for ICD-11. *Psychological Assessment, 30*(2), 154-169. doi:10.1037/pas0000459

Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance, 11*(2-3), 245-269. doi:10.1207/s15327043hup1102&3_7

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. doi:10.1037//0021-9010.81.6.660

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

Pauls, C. A., & Crost, N. W. (2005). Effects of different instructional sets on the construct validity of the NEO-PI-R. *Personality and Individual Differences, 39*, 297-308. doi:10.1016/j.paid.2005.01.003

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

R Core Team (2018). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Re, A. D. (2013). compute.es: Compute Effect Sizes (Version 0.2-2). Retrieved from http://cran.r-project.org/web/packages/compute.es

Revelle, W. (2018). psych: Procedures for Personality and Psychological Research (Version 1.8.4). Evanston, Illinois: Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2*(4), 313-345. doi:10.1111/j.1745-6916.2007.00047.x

Salgado, J. F., & Táuriz, G. (2014). The Five-Factor Model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology, 23*(1), 3-30. doi:10.1080/1359432x.2012.716198

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometric Monograph No.17. doi:http://dx.doi.org/10.1007/BF03372160

Sarkar, D. (2008). *Lattice: Multivariate data visualization with R*. New York: Springer.

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

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*(2), 262-274. doi:10.1037//0033-2909.124.2.262

Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology, 87*(2), 211-219.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*(3), 175-199. doi:10.1207/s15327043hup1903_1

Watrin, L., Geiger, M., Spengler, M., & Wilhelm, O. (2019). Forced-choice versus Likert responses on an occupational Big Five questionnaire. *Journal of Individual Differences*, Advance online publication. doi:10.1027/1614-0001/a000285

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

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

Widiger, T. A., & Samuel, D. B. (2005). Evidence-based assessment of personality disorders. *Psychological Assessment, 17*(3), 278-287. doi:10.1037/1040-3590.17.3.278

Zhang, B., Sun, T., Drasgow, F., Chernyshenko, O. S., Nye, C. D., Stark, S., & White, L. A. (2019). Though forced, still valid: Psychometric equivalence of forced-choice and single-statement measures. *Organizational Research Methods*, Advance online publication. doi:10.1177/1094428119836486

Ziegler, M. (2007). *Situational demand and its impact on construct and criterion validity of a personality questionnaire: State and trait, a couple you just can't study separately!* Dissertation, LMU München: Fakultät für Psychologie und Pädagogik.

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

Table 1

Sample characteristics for the response format groups

Response format group	**N**	**Gender**			**Age**
		% female	**% male**	**% transgender**	***M* (*SD*)**
MFC-matched	593	62.7	36.9	0.3	23.40 (4.09)
MFC-mixed	652	65.0	34.8	0.2	23.39 (4.34)
Rating scale	622	64.7	34.9	0.3	23.64 (4.39)

Note. MFC = multidimensional forced-choice.

Table 2

Assessment of criteria and hypothesized relations to Big Five

Area	Criterion	Question	Response	Expert rating Big Five corr.
Social activities	Facebook user	"I do not have a Facebook account." was provided as an alternative response option to Number of Facebook friends	Checked, not checked	Extraversion
	Number of Facebook friends	"Please enter the number of your friends on Facebook as accurately as possible."	Text box	Extraversion
	Attended parties / month	"On average, how many parties do you go to in a month?"	Text box	Extraversion
	Number dates / month	"On average, how many dates (with the same person or different persons) do you go on in a month?"	Text box	Extraversion
	Number of persons dated / year	"On average, how many persons do you date in a year?"	Text box	Extraversion
Health	Frequency of smoking	"How often do you smoke cigarettes?"	Never, less than once a month, 1 – 3 times a month, 1 – 3 times a week, on 4 -6 days a week, every day	Neuroticism, Conscientiousness
	Number of smoked cigarettes / day	"How many cigarettes do you smoke on average a day?" This question was only presented if a person selected "every day" to the previous question.	Text box	Neuroticism, Conscientiousness
	Frequency of drinking alcohol	"How often do you consume alcoholic beverages?"	Never, ≤ once a month, 2 – 4 times a month, 2 – 3 times a week, ≥ 4 times a week	Extraversion, Conscientiousness
	Life satisfaction	Satisfaction With Life Scale (Diener, Emmons, Larsen, & Griffin, 1985)	7-point scale from *do not agree at all* to *agree completely*	Neuroticism

	Exercise regularly	"Do you exercise regularly (at least once a week)?"	Yes, no	Conscientiousness
Ability	GPA	"What is your current GPA?"	Text box	Conscientiousness
	Intelligence	Numeric module Modularer Kurzintelligenztest (Dantlgraber, 2015)		-
Charity	Donated blood	"Have you ever donated blood?"	Yes, no	Agreeableness
	Charity work	"Do you do any charity work in social organizations?"	Yes, no	Agreeableness
	Voluntary social year	"Did you do a voluntary social year after graduating from high school?"	Yes, no	Agreeableness
Other	Number times traveled abroad > 1 month	"How many times did you travel abroad for longer than one month after graduating from high school?"	Text box	Extraversion, Openness
	Punctuality	Research assistants noted the time participants arrived for the lab session.	0 = unpunctual, 1 = punctual	Conscientiousness
	Job	"Do you have a job?"	Yes, no	Conscientiousness
	Play instrument	"Do you play an instrument?"	Yes, no	Openness
	Engage in extreme sports	"Do you engage in extreme sports such as paragliding?"	Yes, no	Extraversion

Note. Corr. = correlation. The last column indicates which Big Five domain the respective criterion was hypothesized to correlate with > |.15| by the experts.

Table 3

Latent mean differences between honest and fake-good instructions with 95% CI

Format	**Trait**	**Latent mean difference**	***SD***	**Cohen's *d***
MFC-matched	Neuroticism	-0.63 [-0.71; -0.54]	0.82	-0.77 [-0.89; -0.65]
	Extraversion	0.40 [0.32; 0.49]	0.78	0.52 [0.40; 0.63]
	Openness	0.23 [0.13; 0.33]	0.97	0.24 [0.12; 0.35]
	Agreeableness	-0.05 [-0.15; 0.05]	0.59	-0.09 [-0.20; 0.03]
	Conscientiousness	0.47 [0.36; 0.58]	1.17	0.40 [0.29; 0.52]
MFC-mixed	Neuroticism	-0.80 [-0.89; -0.71]	0.84	-0.95 [-1.06; -0.83]
	Extraversion	0.62 [0.54; 0.69]	0.69	0.90 [0.78; 1.01]
	Openness	0.30 [0.20; 0.40]	0.90	0.33 [0.22; 0.44]
	Agreeableness	0.58 [0.47; 0.69]	1.01	0.58 [0.47; 0.69]
	Conscientiousness	0.59 [0.49; 0.68]	0.96	0.61 [0.50; 0.72]
Rating scale	Neuroticism	-2.16 [-2.30; -2.02]	1.56	-1.38 [-1.51; -1.26]
	Extraversion	1.34 [1.23; 1.45]	1.32	1.01 [0.89; 1.13]
	Openness	1.05 [0.94; 1.17]	1.25	0.84 [0.72; 0.96]
	Agreeableness	0.82 [0.71; 0.93]	0.98	0.84 [0.73; 0.96]
	Conscientiousness	2.32 [2.13; 2.50]	1.85	1.25 [1.13; 1.37]

Table 4

Criterion-related validities for the three response format groups under honest and fake-good administrations

Trait	Criterion	MFC-matched		Rating scale		MFC-mixed	
		r **honest**	***r*** **fake-good**	***r*** **honest**	***r*** **fake-good**	***r*** **honest**	***r*** **fake-good**
Neuroticism	Frequency of smoking	0.01	0.08	0.10	0.05	0.06	0.09
	Life satisfaction	-0.44	-0.24	-0.65	-0.21	-0.67	-0.29
Extraversion	Facebook user	0.20	0.18	0.16	0.09	0.22	0.09
	Number of Facebook friends (log)	0.50	0.11	0.43	0.23	0.46	0.19
	Attended parties/month	0.48	0.16	0.49	0.22	0.53	0.21
	Number dates/month	0.13	0.08	0.23	0.13	0.34	0.14
	Number of persons dated/year	0.16	0.02	0.21	0.04	0.35	0.06
	Frequency of drinking alcohol	0.33	0.20	0.29	0.14	0.32	0.10
	Number times traveled abroad > 1 month	0.07	0.14	0.16	0.04	0.24	0.07
	Engage in extreme sports	0.21	0.04	0.21	-0.08	0.19	0.17
Openness	Number times traveled abroad > 1 month	0.15	0.12	0.18	0.02	0.07	-0.04
	Play instrument	-0.03	0.09	0.17	0.14	0.18	0.05
Agreeable-ness	Donated blood	0.07	0.04	0.01	0.00	0.06	0.05
	Charity work	0.04	0.10	0.36	0.20	0.20	0.07
	Voluntary social year	-0.15	0.01	0.08	0.04	0.26	0.06
Conscien-tiousness	Frequency of smoking	-0.19	-0.07	-0.19	-0.07	-0.21	0.05
	Frequency of drinking alcohol	-0.27	0.07	-0.19	0.06	-0.28	0.03
	GPA	-0.14	-0.15	-0.05	-0.12	-0.18	-0.05
	Exercise regularly	0.00	-0.03	0.07	0.06	0.03	0.03
	Punctuality	-0.16	-0.09	0.08	-0.01	0.36	-0.11
	Job	-0.10	0.13	0.04	0.13	-0.06	0.14

Note. Number of Facebook friends was analyzed as a log-transformed variable. The criterion number of cigarettes a day was removed due to estimation problems.

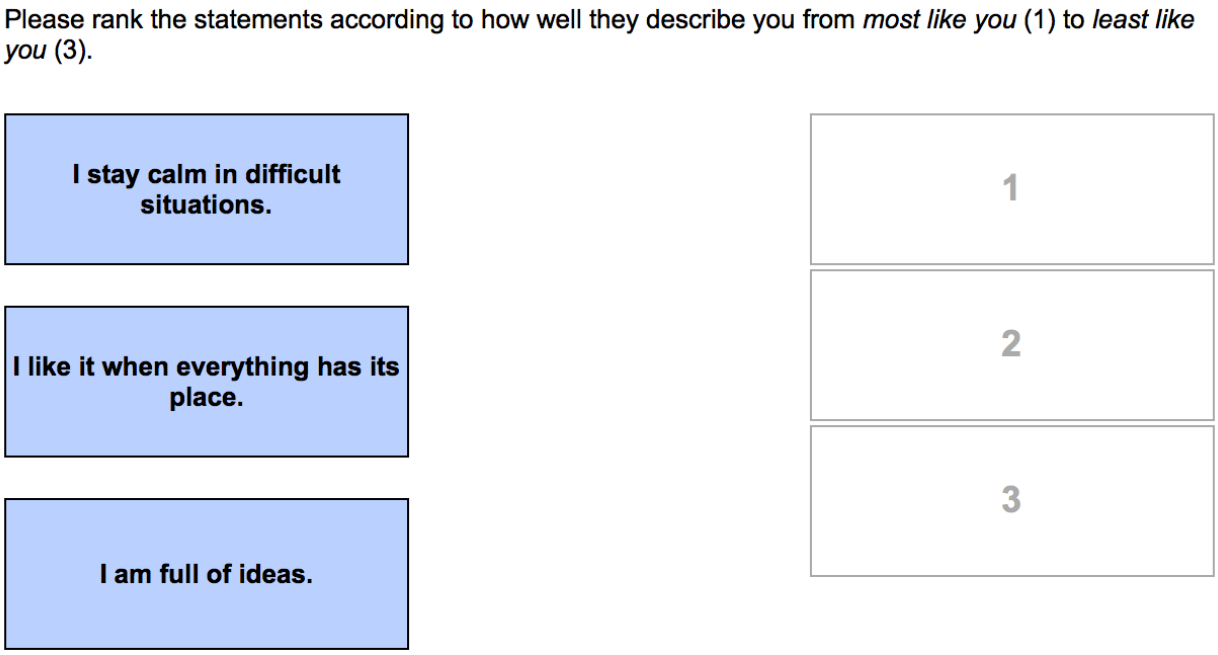

Figure 1. *Sample triplet from the Big Five Triplets*

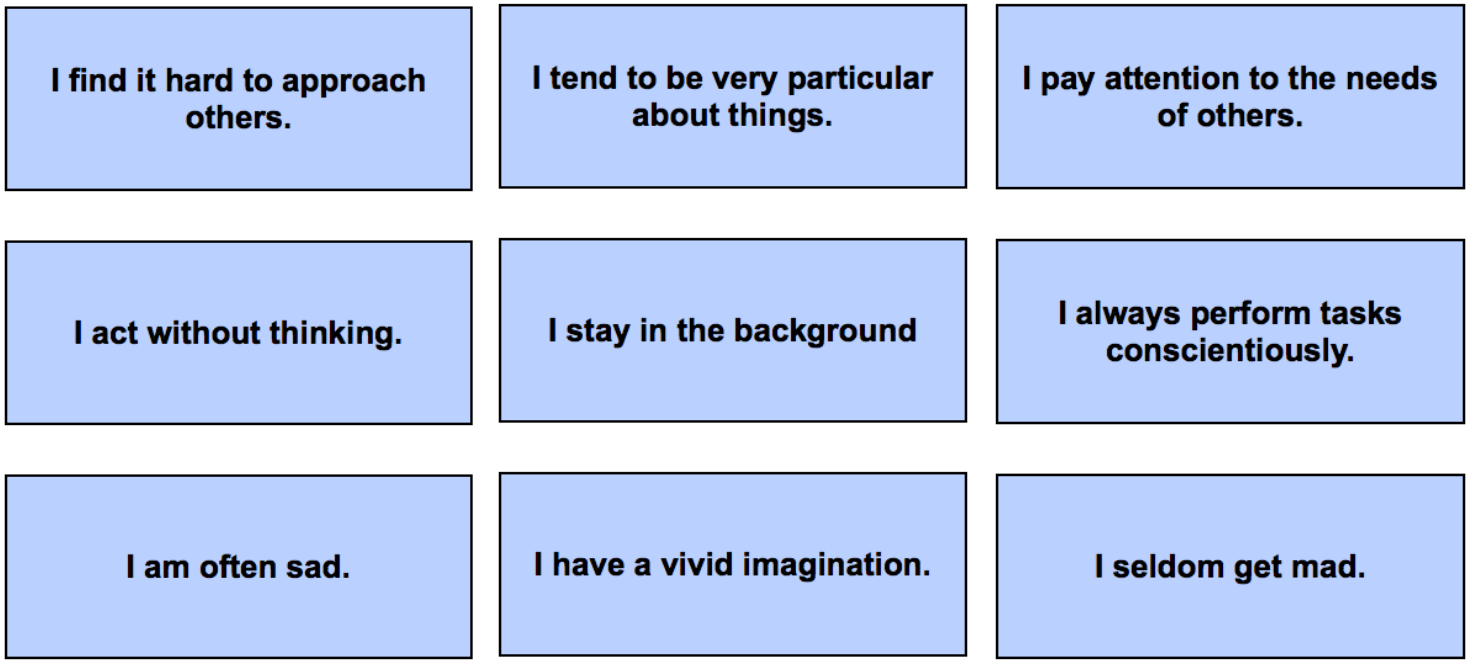

Figure 2. Examples of socially undesirable (left), neutral (middle) and socially desirable (right) triplets from the Big Five Triplets.

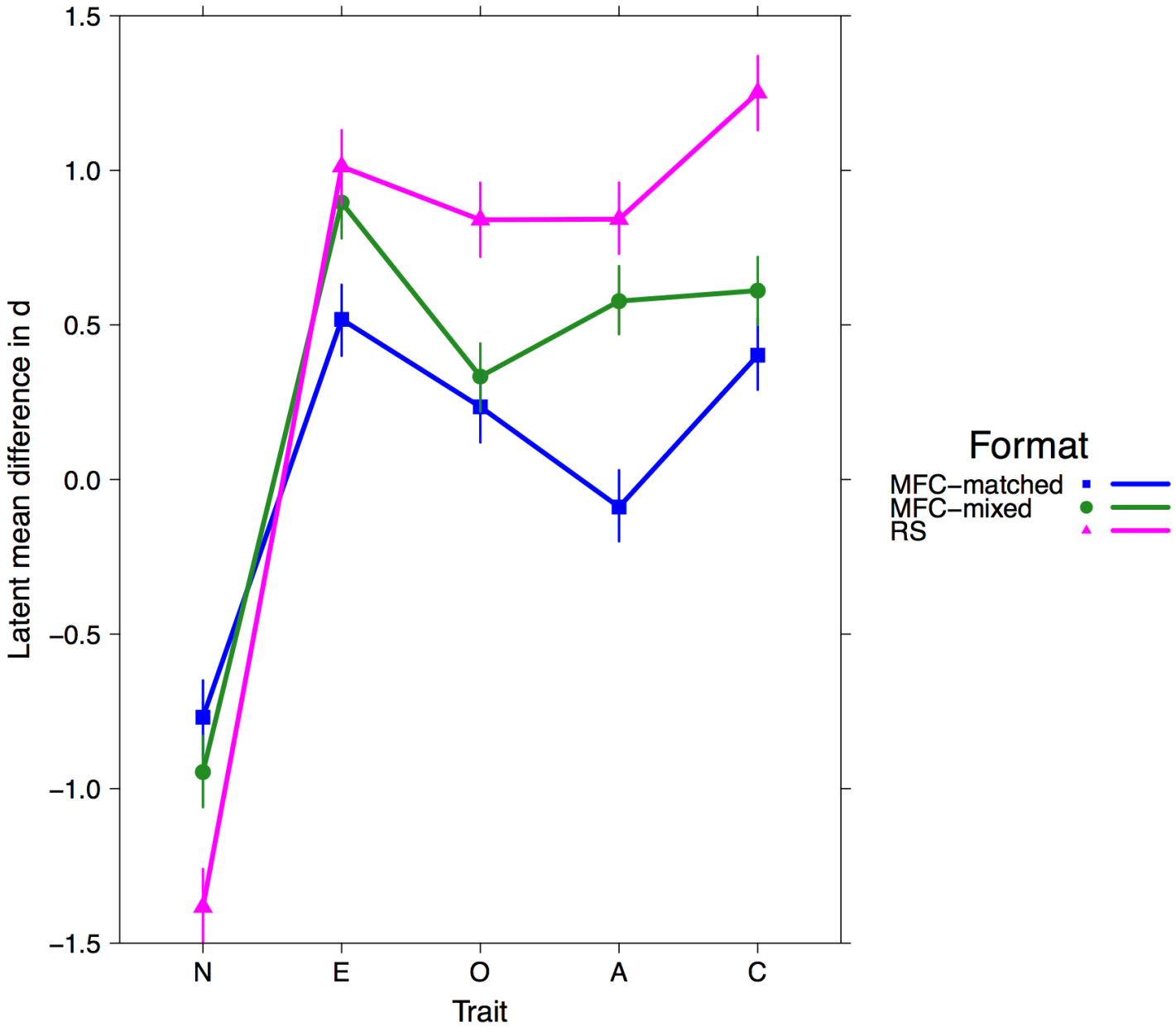

Figure 3. Latent mean differences in Cohen's *d* between the honest and the fake-good instruction for the three response format groups. MFC = multidimensional forced-choice, RS = rating scale, N = neuroticism, E = extraversion, O = openness to experience, A = agreeableness, C = conscientiousness.