Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

The Psychometric Costs of Applicants' Faking: Examining Measurement Invariance and Retest Correlations Across Response Conditions

Georg Krammer, Markus Sommer & Martin E. Arendasy

To cite this article: Georg Krammer, Markus Sommer & Martin E. Arendasy (2017) The Psychometric Costs of Applicants' Faking: Examining Measurement Invariance and Retest Correlations Across Response Conditions, Journal of Personality Assessment, 99:5, 510-523, DOI: 10.1080/00223891.2017.1285781

To link to this article: https://doi.org/10.1080/00223891.2017.1285781

JOURNAL OF PERSONALITY ASSESSMENT
2017, VOL. 99, NO. 5, 510–523
https://doi.org/10.1080/00223891.2017.1285781

Routledge
Taylor & Francis Group

ARTICLES

The Psychometric Costs of Applicants' Faking: Examining Measurement Invariance and Retest Correlations Across Response Conditions

Georg Krammer,[1] Markus Sommer,[2] and Martin E. Arendasy[2]

[1]Institute of Practical Education and Practitioner Research, University College of Teacher Education Styria, Graz, Austria; [2]Psychological Methodology and Diagnostic, Department of Psychology, University of Graz, Austria

ABSTRACT

This study examines the stability of the response process and the rank-order of respondents responding to 3 personality scales in 4 different response conditions. Applicants to the University College of Teacher Education Styria ($N = 243$) completed personality scales as part of their college admission process. Half a year later, they retook the same personality scales in 1 of 3 randomly assigned experimental response conditions: honest, faking-good, or reproduce. Longitudinal means and covariance structure analyses showed that applicants' response processes could be partially reproduced after half a year, and respondents seemed to rely on an honest response behavior as a frame of reference. Additionally, applicants' faking behavior and instructed faking (faking-good) caused differences in the latent retest correlations and consistently affected measurement properties. The varying latent retest correlations indicated that faking can distort respondents' rank-order and thus the fairness of subsequent selection decisions, depending on the kind of faking behavior. Instructed faking (faking-good) even affected weak measurement invariance, whereas applicants' faking behavior did not. Consequently, correlations with personality scales—which can be utilized for predictive validity—may be readily interpreted for applicants. Faking behavior also introduced a uniform bias, implying that the classically observed mean raw score differences may not be readily interpreted.

ARTICLE HISTORY
Received 2 May 2016
Revised 4 December 2016

Faking is a type of response bias wherein respondents distort their responses to personality scale items to be viewed more favorably (McFarland & Ryan, 2000). Several studies have indicated that faking can affect the psychometric properties of personality scales (e.g., Hartman & Grubb, 2011; Miller & Ruggs, 2014; Zickar & Robie, 1999; Ziegler & Bühner, 2009). Furthermore, selection decisions based on personality scales might be detrimentally affected due to applicants differing in their propensity and intensity to fake (e.g., Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Mueller-Hanson, Heggestad, & Thornton, 2003; Rosse, Stecher, Miller, & Levin, 1998; Winkelspecht, Lewis, & Thomas, 2006; Zickar & Robie, 1999). Therefore, it is important to examine how applicants fake personality scales in real-life selection settings (Goffin & Boyd, 2009; Griffith & Peterson, 2011; Kuncel, Goldberg, & Kiger, 2011). To this end, this study examined four different response behaviors to personality scales: (a) a real-life admission testing situation, (b) a classic honest condition, (c) an instructed faking-good condition, and (d) a condition prompting incumbents to reproduce the response behavior from their admission testing setting. The comparison of the psychometric characteristics of the personality scales across these four response conditions is expected to shed some light on the processes used by applicants to fake personality scales in real-life selection settings.

Effects of faking on personality scale scores

Practitioners are mainly concerned about the effects of faking on applicants' personality scale scores, and the effects on subsequent selection decisions. The high stakes involved in some personnel selection and educational admission settings are likely to prompt some sort of faking behavior from at least some applicants (cf. Dilchert, Ones, Viswesvaran, & Deller, 2006; Tett & Simonet, 2011). Several laboratory studies using an instructed faking-good condition (i.e., instructions to fake good) indicate that respondents are capable of increasing their personality scale scores ($0.48 \leq d \leq 0.65$ in within-subject designs, and $0.47 \leq d \leq 0.93$ in between-subject designs: Viswesvaran & Ones, 1999). However, some scholars argue that these detrimental effects of faking are specific to laboratory studies, and that in real-life selection settings, effects of faking on applicants' personality scale scores are negligible (e.g., Bradley & Hauenstein, 2006; Hogan, Barrett, & Hogan, 2007; Ones, Dilchert, Viswesvaran, & Judge, 2007). This view has been partially supported by a recent meta-analysis indicating only negligible to small differences between applicants and incumbents ($0.11 \leq d \leq 0.45$; Birkeland et al., 2006). Unfortunately, the results of the meta-analysis just cited are necessarily based on the assumption that strong measurement invariance is given

CONTACT Georg Krammer georg.krammer@phst.at Institute of Practical Education and Practitioner Research, University College of Teacher Education Styria, Hasnerplatz 12, 8010 Graz, Austria.

across applicants and incumbents. If strong measurement invariance is given, respondents with the same standing on the latent trait(s) have equal expected item scores, test scores, or both. Thus, mean differences within and between groups of respondents (e.g., applicants vs. incumbents) can be interpreted at face value (e.g., Millsap, 2011; Mislevy et al., 2013). By contrast, if strong measurement invariance is violated, mean score differences reflect not only (latent) true score differences between the groups, conditions, or both, but additionally measurement bias (cf. Li & Zumbo, 2009).

Studies examining retest correlations of personality scales completed under different response conditions offer some tentative evidence that faking might affect the psychometric properties of personality scales. For example, Peterson, Griffith, Isaacson, O'Connell, and Mangos (2011) retested applicants as incumbents, and reported a retest correlation for conscientiousness of $r = .62$. Similar results were obtained by Griffith, Chmielowski, and Yoshita (2007), who reported a retest correlation for conscientiousness of $r = .50$ for applicants responding to a conscientiousness scale with and without motivation to fake. In line with these findings, Hogan et al. (2007) reported retest correlations for the Big Five ranging from $r = .46$ to $r = .68$. In contrast to these findings, merely retesting volunteers generally yields higher retest correlations (e.g., $.70 \leq r \leq .90$: Arendasy, Sommer, & Feldhammer, 2011; $.67 \leq r \leq .85$: McCrae, Yik, Trapnell, Bond, & Paulhus, 1998; $.50 \leq r \leq .85$: Hogan, 1992), as has already been pointed out by scholars (e.g., Hogan et al., 2007). These lower than expected retest correlations for applicants compared to volunteers might indicate that faking compromises the measurement invariance of personality scales.

Factors affecting faking on personality scale items

To interpret personality scale scores at face value, individual differences in the respondents' personality scale scores should be entirely attributable to individual differences in the respondents' standing on the latent trait measured (e.g., Millsap, 2011; Mislevy et al., 2013). This implies that respondents' standing on the latent trait should be the only individual differences construct having a causal effect on their responses to the personality scale items. However, several studies indicate that applicants consider a wide range of aspects when deciding which response category to endorse on personality scale items (e.g., Donovan, Dwight, & Hurtz, 2003; König, Merz, & Trauffer, 2012; Kuncel & Tellegen, 2009; Ziegler, 2011). For instance, applicants evaluate the relevance of individual items for a particular position or education when considering faking their responses (cf. König et al., 2012; Ziegler, 2011). Thus, their responses do not only depend anymore on their standing on the latent trait, but also on their perceived relevance of the items, and their willingness to fake.

Several theoretical models have been proposed to explain why and how applicants fake in real-life selection settings (e.g., Goffin & Boyd, 2009; Marcus, 2009; McFarland & Ryan, 2000, 2006; Mueller-Hanson, Heggestad, & Thornton, 2006; Snell, Sydell, & Lueke, 1999). Despite various differences, all theoretical models just cited distinguish between dispositional and situational antecedents of applicants' faking behavior. Dispositional antecedents refer to individual differences (e.g., regarding honesty) or to appraisals of the relevance of personality traits for a particular job. On the other hand, situational antecedents refer to more fluctuating characteristics of the assessment situation, such as the perceived attractiveness of the job or the perceived competition level. In principle, these theoretical models can also be applied to an instructed faking-good condition in a laboratory. For instance, when instructed to make a favorable impression for a particular job, one might expect that respondents consider the perceived relevance of personality scale items for this particular job. However, individual differences in the accuracy of the perceived relevance might be larger in the laboratory than in real-life selection settings. By contrast, individual differences in respondents' willingness to fake and its antecedents might be less pronounced in laboratory settings than in a real-life selection setting. This raises the question of whether personality scales can be assumed to exhibit measurement invariance across various response conditions (faking good, honest, applicants, etc.).

Effects of faking on measurement properties and test scores

Researchers have found it useful to distinguish four levels of measurement invariance (cf. Cheung & Rensvold, 2002; Millsap, 2011; Mislevy et al., 2013; Raykov, Marcoulides, & Li, 2012; Vandenberg & Lance, 2000). Each level builds on the prior level, and they are tested level by level. The lowest level is *configural measurement invariance*. The second level is *weak measurement invariance*. The third level is *strong measurement invariance*. The final level is *strict measurement invariance*. Faking behavior can theoretically affect personality scales on each level of measurement invariance, and personality test scores, which will be outlined in the following.

Effects of faking on the levels of measurement invariance

Configural measurement invariance only assumes the general factor structure of personality scales to be equal across response conditions. Configural measurement invariance would be violated when a personality scale measures one latent trait in one response condition, but more than one latent trait in another response condition. For example, consider a questionnaire assessing the Big Five domain factor Extraversion. The questionnaire might measure extraversion for incumbents. However, for applicants it might measure extraversion and faking intensity. In this case, configural measurement invariance would be violated. Consequently, scores of incumbents and applicants could not be directly compared in a meaningful way (e.g., Cheung & Rensvold, 2002; Millsap, 2011; Mislevy et al., 2013; Vandenberg & Lance, 2000).

Weak measurement invariance additionally assumes the items to be equally saturated by the latent trait they have been intended to measure across response conditions. Weak measurement invariance would be violated when an item reflects its latent trait to a different extent across response conditions. For example, consider the extraversion questionnaire again and an applicant for a teacher vacancy. Now, let one item be "I like to stay in contact with parents." Applicants might appraise this item as an indicator of their extraversion. For incumbents by

contrast, this item content might be a necessary part of their job, and thus the item might be only moderately related to their extraversion. In this example, the item would reflect individual differences in extraversion for applicants, but would do so less in a sample of incumbents. The violation of weak measurement invariance also indicates that a nonuniform bias is present (for a detailed discussion of nonuniform bias cf. Barendse, Oort, & Garst, 2010; Penfield & Camilli, 2007). Furthermore, weak measurement invariance across response conditions is necessary to compare correlations within one response condition to correlations within another response condition (Chen, 2008).

Strong measurement invariance additionally assumes the baseline probability of endorsing a specific response category to be equal across response conditions. Strong measurement invariance would be violated when respondents with the same standing on a latent trait have unequal item scores across different response conditions. Consider a respondent who moderately likes "staying in contact with parents." As an applicant for a teacher vacancy however, he or she might indicate to like "staying in contact with parents" to a high degree. With faking-good instructions for a teacher vacancy, he or she might even state to extremely like "staying in contact with parents." In this example, strong measurement invariance would be violated. This level of measurement invariance is necessary to interpret mean score differences at face value. If it is not given, a uniform bias is present (for a detailed discussion of uniform bias, cf. Barendse et al., 2010; Penfield & Camilli, 2007). This implies that mean score differences do not only reflect mean individual differences, but also measurement bias (cf. Chen, 2008; Cheung & Rensvold, 2002; Li & Zumbo, 2009; Millsap, 2011).

Strict measurement invariance finally also assumes both the systematic variance not explained by the psychometric model and the unsystematic variance of the measurement error to be equal across response conditions. Strict measurement invariance would be violated when responses are affected by their latent trait and situational factors. For example, applicants might also consider how badly they need a job when responding to items. This would introduce systematic variance in the item responses unaccounted for by the psychometric model. An instruction to fake good could reverse this effect again, as all respondents might respond as if they desperately needed the job.

Effects of faking on latent means and variance estimates

Faking might affect the levels of measurement invariance, but also the latent means and variances of personality scales. When responding as applicants trying to qualify for a position, respondents might try to elevate their scores by endorsing higher response categories. If this is done equally for all items of a scale, then an increase in the latent mean will result. However, when raising scores, initially lower scores can be inflated by a greater extent, so respondents with a lower standing on the latent trait can raise their scores to a greater extent than respondents with an already higher standing on the latent trait (e.g., Goffin & Boyd, 2009; McFarland & Ryan, 2000, 2006). This in turn can lead to a decrease of the variance of the individual differences construct measured.

The extent to which faking affects the latent trait means and variances of a personality scale also depends on how homogeneously faking affects the item responses. Consider, for example, items of a personality scale differing in their perceived relevance for a position. If respondents choose to fake, the extent to which they will increase their scores will differ across these items. This in turn will lead to a violation of strong measurement invariance due to a uniform measurement bias (cf. Barendse et al., 2010; Penfield & Camilli, 2007). In contrast, if the responses to all items of one scale are elevated by the same extent, faking would increase the latent trait mean. Such a response behavior could not be distinguished from truly having a higher standing on the latent trait. Note that these two possible outcomes are not mutually exclusive. For instance, respondents might perceive all items of a personality scale as being relevant for a position, and still perceive selected items as being particularly relevant. In this case, the baseline probabilities of endorsing a specific response category would be globally higher, and even higher for some items. This global increase would generalize to the latent mean, and the additional increase for selected items would violate strong measurement invariance. As a consequence, one would expect to observe a shift of the latent mean and a lack of strong measurement invariance.

In line with these considerations, studies examining the level of measurement invariance across honest and faking-good conditions yield results ranging from configural measurement invariance across these two conditions (e.g., Miller & Ruggs, 2014), to weak measurement invariance (e.g., Zickar & Robie, 1999), up to strong measurement invariance (Ferrando & Anguiano-Carrasco, 2009a, 2009b). The inconsistent findings in these studies could be attributable to (a) the specific personality constructs measured and their perceived relevance for the position (e.g., psychoticism: Ferrando & Anguiano-Carrasco, 2009a; impression management: Miller & Ruggs, 2014), (b) the samples of respondents used in these studies (e.g., military: Zickar & Robie, 1999; undergraduate business class students: Miller & Ruggs, 2014), or (c) differences in test design characteristics such as item presentation mode (e.g., blocked: Ferrando & Anguiano-Carrasco, 2009a; random: Zickar & Robie, 1999).

Unfortunately, studies examining measurement invariance across real-life selection settings and other response conditions are sparse. Robie, Zickar, and Schmit (2001) examined measurement invariance across applicants for a sales manager position and job incumbents. The authors reported strict measurement invariance for five out of six personality scales examined. The one scale for which strict measurement invariance was violated exhibited a measurement bias for two items: One item favored applicants and one item favored incumbents. Based on these mixed results, the authors argued for bias cancelation. Taken together, the currently available empirical evidence suggests that faking might affect the level of measurement invariance of personality scales. However, the extent to which this is the case seems to depend on the types of faking behavior compared (i.e., faking-good or applicants' faking), the position in question, and the characteristics of the personality scales used.

Stability of applicants' faking behavior

Although several studies examined measurement invariance and mean raw score differences across various response

conditions, little is known about the stability of individual differences in applicants' response behavior. Stability of response behavior has two important aspects that need to be distinguished: (a) the stability of the response process, and (b) the stability of the rank-order of the respondents. The former can be addressed by examining measurement invariance, whereas the latter can be addressed by retest correlations. Both aspects only partially build on each other. When configural or weak measurement invariance have been violated across response conditions, estimates of the retest correlations within response conditions might be biased and therefore cannot be directly compared to each other. By contrast, when weak measurement invariance across response conditions holds, retest correlations can be readily interpreted and compared across pairs of response conditions (Chen, 2008). Unfortunately, neither weak nor strong measurement invariance guarantees that the rank-order of respondents is preserved across response conditions. Measurement invariance merely indicates the extent to which the response process can be assumed to be invariant across response conditions (i.e., stability of the response process). For instance, Zickar and Robbie (1999) outlined that faking might only result in a shift of the latent trait without affecting strong or strict measurement invariance. According to this model, applicants should fake personality scales in selection settings by imagining a more favorable version of themselves. If this model has merit, one would expect the response process to the personality scales to be identical across, for example, honest responding and real-life selection settings. However, because respondents differ in their propensity and intensity to fake (e.g., Birkeland et al., 2006; Mueller-Hanson et al., 2003; Rosse et al., 1998; Winkelspecht et al., 2006; Zickar & Robie, 1999), the retest correlations of personality scales across these two response conditions might still be detrimentally affected. This could explain why retesting applicants generally yields smaller retest correlations than retesting volunteers (cf. Arendasy et al., 2011; Griffin & Wilson, 2012; Griffith et al., 2007; Hogan et al., 2007; Hogan, 1992; McCrae et al., 1998; Peterson et al., 2011). This could also question the interpretation of test scores obtained from incumbents who were instructed to respond like applicants (e.g., Fell & König, 2016). Therefore, it is important to examine both aspects of stability separately. Unfortunately, this has not been done so far.

Description of this study

This study examined the level of measurement invariance of three personality scales in a real-life admission testing situation across four different response conditions. This was done in a combined within-subject and between-subject design. The research design is illustrated in Figure 1. At the first time point of measurement (selection, t1), respondents completed the personality scales as part of their admission testing process for studying at the University College of Teacher Education Styria. The personality scales of interest were one part of their admission process. Half a year later (retest, t2), the personality scales were given for a second time to the applicants, who were by now incumbents. The incumbents were randomly assigned to one of the three response conditions, which only differed in the instructions of the personality scales. The first

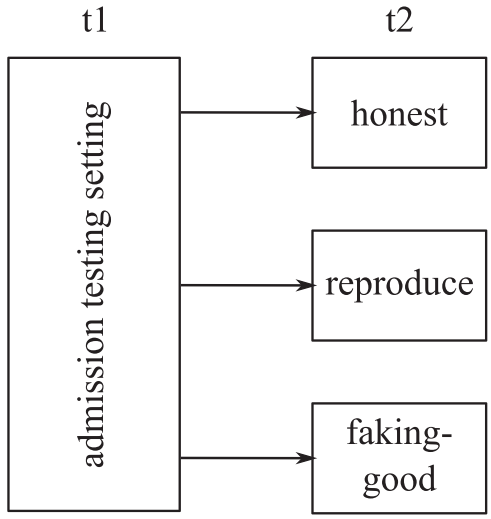

Figure 1. The combined within-subject and between-subject design. At selection (t1), all respondents were applicants and went through the same admission testing situation. At retest (t2), the same applicants were by now incumbents, and were randomly assigned to one of three response conditions.

response condition was the *honest condition*, for which incumbents were given the personality scales with the standard instructions. The second response condition was the *reproduce condition*, for which incumbents were prompted to recall the situation of their selection process and to respond to the items as they responded when they first took the admission test. The third response condition was the *faking-good condition*, for which incumbents were instructed to portray themselves as favorably as possible to be admitted to the University College of Teacher Education Styria. This specific research design was chosen because a combination of a real-life selection setting and a within-subject design has been argued to constitute the gold standard in faking research (e.g., Komar, Brown, Komar, & Robie, 2008; Peterson et al., 2011; Ryan & Boyce, 2006). Furthermore, the research design allowed us to examine both aspects of stability.

Research hypotheses

For respondents in the honest condition, we examined measurement invariance across the real-life admission testing setting and the laboratory honest responses. Scholars (e.g., Bradley & Hauenstein, 2006; Hogan et al., 2007; Ones et al., 2007) claimed that applicants' faking behavior only negligibly affects their personality test scores. In line with this hypothesis, strict measurement invariance was found across applicants and incumbents (Robie et al., 2001). Therefore, we hypothesized that we would find strict measurement invariance across applicants and incumbents instructed to respond honestly (Hypothesis 1). However, previous studies indicated that faking distorts applicants' rank order (e.g., Mueller-Hanson et al., 2003; Rosse et al., 1998; Winkelspecht et al., 2006). Consistent with this hypothesis, moderate retest correlations were found between applicants and incumbents (e.g., Griffin & Wilson, 2012; Griffith et al., 2007; Peterson et al., 2011). We therefore hypothesized that we would observe moderate retest correlations between the real-life admission testing setting and the honest responses (Hypothesis 2). Finally, samples of applicants have

exhibited higher mean scores than samples of incumbents, although these effects were usually small (Birkeland et al., 2006). Therefore, we also hypothesized higher latent means of the applicants compared to the incumbents in our study (Hypothesis 3).

The faking-good condition compared applicants' response behavior in the real-life admission testing situation to the response behavior in an instructed faking-good condition. Instructed faking differs from applicants' faking, as all respondents can be assumed to fake irrespectively of individual differences in faking propensity (cf. Robie et al., 2001). Based on previous studies, we expected to observe at least configural measurement invariance (cf. Miller & Ruggs, 2014). Because some studies (cf. Ferrando & Anguiano-Carrasco, 2009a, 2009b) also observed strong measurement invariance, we hypothesized that it might even be possible that strong measurement invariance holds across these two response conditions (Hypothesis 4). Furthermore, we hypothesized that we would find rather low retest correlations between these two response conditions, because the lack of individual differences in faking propensity in the instructed faking-good condition can be assumed to alter the rank-order of the respondents (Hypothesis 5). This prediction is also in line with previous studies examining the retest correlations between applicants and incumbents or volunteers instructed to fake good (cf. Ellingson, Sackett, & Hough, 1999; Griffith et al., 2007). For the same reason, instructed faking-good should also lead to an increase in latent means compared to the applicants (Hypothesis 6).

The last condition examined the stability of applicants' response behavior across a real-life admission testing situation and the reproduced response behavior (reproduce condition). Based on the finding that the intensity and propensity of applicants' faking did not increase when rejected applicants reapplied for the same job (Hogan et al., 2007), we hypothesized applicants' faking behavior to be relatively stable. On the one hand, this should be reflected in the stability of the response process and would imply strict measurement invariance for incumbents attempting to reproduce their applicants' response behavior (Hypothesis 7). On the other hand, the retest correlation should also be high (Hypothesis 8). Finally, we expected that respondents should be able to reproduce the intensity of their faking behavior. As a consequence, latent mean differences across these two response conditions should be negligible (Hypothesis 9).

Method

Measures

Studying applicants' faking behavior requires a personality inventory that is used for selection purposes. The Inventory for Personality Assessment in Situations (IPS; Schaarschmidt & Fischer, 2013) is such a personality inventory, which is used for admission testing processes in teacher education (e.g., Schulz-Kolland, Krammer, Rottensteiner, & Weitlaner, 2014). The IPS is suitable for professions with high psychosocial demands, such as the teaching profession. Validation studies have shown that the IPS measures personality dispositions relevant to teaching training and to the teaching profession, and that there are characteristic personality profiles of both teachers and students in teacher education, which differ markedly from students of other courses of study (cf. Krammer, Sommer, & Arendasy, 2016; Mayr & Brandstätter, 1998; Schaarschmidt & Fischer, 2013).

The IPS is comprised of three higher order factors: *social and communicative behavior*, *health and recreational behavior*, and *performance behavior*. To study applicants' faking behavior, one subscale of each higher order factor was chosen. *Activity in familiar communicative situation* was chosen for the higher order factor social and communicative behavior, *preventive health behavior in response to warning signals* for the higher order factor health and recreational behavior, and *self-confidence in test situation* for the higher order factor performance behavior. Previous research demonstrates that activity in familiar communicative situation is highly related to the Big Five domain factor Extraversion, and self-confidence in test situation is highly related to the Big Five domain factor Emotional Stability/Neuroticism (cf. Koschmieder, Pretsch, & Neubauer, 2015). This could be due to the fact that the item contents of the scales represent sociableness and carefreeness, respectively, which are both highly saturated by their respective domain factors Extraversion and Emotional Stability (Arendasy et al., 2011; Costa & McCrae, 1992). Extraversion and Emotional Stability in turn have been shown to be relevant for success in teacher education and the teaching profession (e.g., Mayr, 2011). Finally, preventive health behavior in response to warning signals is relevant for teacher education and the teaching profession, as teachers are a high-risk group for burnout (e.g., Unterbrink et al., 2007), and the pressure of the profession can manifest itself already during education (cf. Gold & Roth, 1993).

All IPS scales consisted of a situational context and five personal behaviors that might occur in the described situation. Respondents had to indicate the extent to which they would exhibit each personal behavior in the described situation using a 4-point Likert scale ranging from *not true at all* (1) to *definitely true* (4). The situational context for activity in familiar communicative situation describes the setting of a sociable group of friends and acquaintances, and the items relate to the activity in this situation (e.g., "be very lively"). For the scale preventive health behavior in response to warning signals, the situational context is beginning to feel incapacitated, people already noticing this, and people suggesting a visit to the doctor. The items in this situation relate to the extent to which these warning signals are heeded (e.g., "actively do something for my health"). Finally, the scale self-confidence in test situation describes a test situation and emphasizes the emotional burden of having to pass the test situation successfully. The items in this situation relate to one's own reaction to this pressure (e.g., "stay calm"). To facilitate the interpretation, all items were recoded so that higher scores were more favorable. All three scales showed sufficient internal consistency at selection (t1: $.639 \leq \alpha \leq .757$). At retest (t2), all three scales exhibited good internal consistency in the honest condition and reproduce condition ($.720 \leq \alpha \leq .875$), whereas the internal consistencies differed in the faking-good condition ($\alpha = .529$, $\alpha = .747$, and $\alpha = .862$, respectively).

Procedure

The research design is illustrated in Figure 1. The completion of personality scales for selection (t1) was a part of the admission

testing process of the University College of Teacher Education Styria (for further details on the admission process, see Schulz-Kolland et al., 2014). For the admission testing process, the applicants had to demonstrate adequate vocal function with no vocal chord disorders, followed by a test of their German skills, the measures for intelligence and personality, and finally an interview. All subtests of the admission testing process were relevant for admission. Applicants were ranked according to their final score, which was a weighted linear combination of all subtests. Applicants were aware that every part of the admission testing process would count for their ranking, but were unaware of the respective weights. Typically, approximately 60% of the applicants are admitted every academic year.

At retest (t2), the applicants were retested as incumbents with the same personality scales. The incumbents were randomly assigned to the three experimental response conditions. In each experimental response condition (honest, reproduce, or faking-good) the personality scales were completed with differing instructions. Because of the instructions, respondents in the reproduce and faking-good conditions were necessarily aware of the fact that the study was concerned with faking behavior. In contrast, respondents in the honest condition were only given the standard instructions of the personality scales to collect responses that were as honest as possible. These respondents were debriefed afterward. The data of selection (t1) and of retest (t2) were matched by a code, which could be generated out of the data at selection (t1) and that respondents were asked to generate themselves at retest (t2).

Sample

A total of 175 (72.02%) females and 68 (27.98%) males aged 18 to 44 years ($M = 23.2$, $SD = 4.8$) participated in this study. At retest (t2), all respondents were in their second semester of a bachelor's degree to become either primary ($n = 131$, $\sim 53.9\%$) or secondary ($n = 112$, $\sim 46.1\%$) school teachers. Even though the respondents were randomly assigned to the three response conditions, the comparability of the three subsamples was examined. The respondents of the three response conditions ($n = 81$ each) did not differ with regard to gender ($\chi^2[2] = 2.001$, $p = .368$, $V = .091$); age ($F[2, 240] = 2.932$, $p = .055$, partial $\eta^2 = .024$); and their study majors ($\chi^2[2] = 0.629$, $p = .730$, $V = .051$). The data were collected over a time span of 2 years. Therefore, the comparability across the 2 years of data collection was examined. There were no differences in gender ($\chi^2[1] = 0.067$, $p = .796$, $V = .017$); age ($t[241] = 1.226$, $p = .222$, $d = 0.158$); and study major ($\chi^2[1] = 0.141$, $p = .707$, $V = .024$), across the 2 years. The data obtained in the first year of data collection had been used in a previous study examining the prevalence of faking and its effect on the raw scores of several personality scales (Krammer & Pflanzl, 2015).

Tested models

Measurement invariance, latent retest correlations, and latent mean shifts were analyzed using multigroup means and covariance structure analysis. The multigroup means and covariance structure model is depicted in Figure 2. The three response conditions constituted the three groups. For each response condition, a one-factor model was specified for each measurement point. The latent factor at selection (t1) was hypothesized to correlate with the latent factor at retest (t2). The unstandardized factor loading of the first item was set to 1 for identification purposes at both measurement points. To account for the readministration of the same items, the residuals of the respectively same items were allowed to correlate. First, configural measurement invariance was tested, by specifying the described two-factor model in all three response conditions. Next, equality constraints were imposed to (a) test the comparability of the three response conditions at selection (t1), and (b) test measurement invariance across the measurement points. The model fit was evaluated using the following goodness-of-fit statistics: nonsignificant χ^2 statistic, root mean square error of approximation (RMSEA) $< .06$, and comparative fit index (CFI) $\geq .95$ (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004).[1] Each model fit was also compared to its less restrictive precursor using the following criteria: nonsignificant $\Delta\chi^2$ statistic and ΔCFI $\leq .002$ (Meade, Johnson, & Braddy, 2008). Because univariate (skewness and kurtosis $> |1|$) and multivariate (multivariate skewness > 26.668, all $p < .001$; MVN package; Korkmaz, Goksuluk, & Zararsiz, 2014) normality were violated, the parameters were estimated using a maximum likelihood estimator with robust standard errors and a Satorra–Bentler scaled test statistic (MLM; Satorra & Bentler, 1994). This parameter estimation method has been shown to perform well under similar conditions in simulation studies (e.g., Curran, West, & Finch, 1996). For small sample sizes, this estimator also outperforms the means and variance-adjusted weighted least squares estimator (WLSMV) for measurement invariance analysis (cf. Sass, Schmitt, & Marsh, 2014). All calculations were carried out in R (R Core Team, 2015) using the lavaan package (Rosseel, 2012).

Ideally, the admission testing situation should exhibit strict measurement invariance, and equal latent means and variances across the three response conditions. To test these hypotheses, we first constrained the selection (t1) factor loadings to be equal across the three conditions ($\lambda_{t1,h,1} = \lambda_{t1,r,1} = \lambda_{t1,f,1}$, etc.) to test for weak measurement invariance. Then, the selection (t1) item intercepts were constrained to be equal across the three conditions ($\tau_{t1,h,1} = \tau_{t1,r,1} = \tau_{t1,f,1}$, etc.) to test for strong measurement invariance. Afterward, the selection (t1) residual variances were constrained to be equal across the three conditions ($\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t1,r,1} = \theta\varepsilon_{t1,f,1}$, etc.) to test for strict measurement invariance. Finally, the latent trait variances and means were constrained to be equal across the three conditions ($\Phi_{t1,h} = \Phi_{t1,r} = \Phi_{t1,f}$ and $\xi_{t1,h} = \xi_{t1,r} = \xi_{t1,f}$, respectively).

After testing the comparability of the response conditions at selection (t1), measurement invariance across the measurement points was examined. First, the factor loadings were

[1] The widely used standardized root mean square residual (SRMR) was omitted in our study. Studies suggest the SRMR is influenced by the sample size, and advise against using ΔSRMR for measurement invariance analysis (e.g., Meade et al., 2008). In accordance with these studies, Monte Carlo simulation studies on our final models suggested the SRMR to be heavily biased in evaluating our model fits, with the sample size causing Type I errors of 99% to 100%. A detailed summary of the Monte Carlo simulation studies is available from the first author on request.

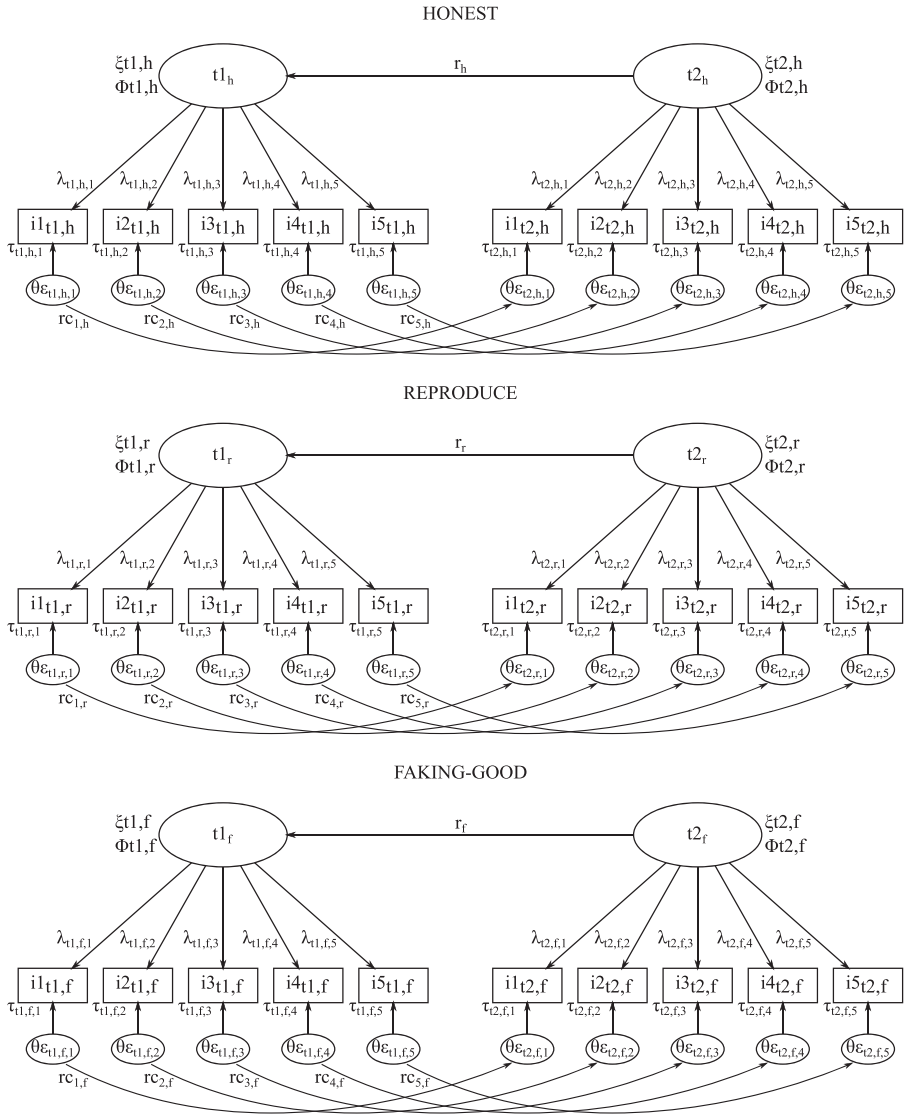

Figure 2. The multigroup means and covariance structure model, containing one-factor models at selection (t1) and at retest (t2), and the covariances. The λs denote the factor loadings, the τs the item intercept, the $\theta\varepsilon$s the residual variances, the *rc*s the residual covariances, the ξs the latent trait means, the Φs the latent trait variances, and *r*s the latent trait covariance.

constrained to be equal across the measurement points in each response condition to test for weak measurement invariance (i.e., $\lambda_{t1,h,1} = \lambda_{t2,h,1}$, $\lambda_{t1,h,2} = \lambda_{t2,h,2}$, etc.). Second, the item intercepts were constrained to be equal across the measurement points in each response conditions to test for strong measurement invariance (i.e., $\tau_{t1,h,1} = \tau_{t2,h,1}$, $\tau_{t1,h,2} = \tau_{t2,h,2}$, etc.). Third, the residual variances were constrained to be equal across the measurement points in each response conditions to test for strict measurement invariance (i.e., $\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t2,h,1}$, $\theta\varepsilon_{t1,h,2} = \theta\varepsilon_{t2,h,2}$, etc.). Afterward, the equality of the residual covariances across the response conditions ($rc_{1,h} = rc_{1,r} = rc_{1,f}$, etc.) was tested. Finally, the equality of the latent trait variances ($\Phi_{t1,h} = \Phi_{t2,h}$, etc.) and means ($\xi_{t1,h} = \xi_{t2,h}$, etc.) across the measurement points was tested.

If any of these equality constraints failed to fit the data, constraints were successively relaxed, and the models again compared (cf. Meredith, 1993; Vandenberg & Lance, 2000). This relaxation of parameters lead to partial invariance models. The partial invariance models were then compared to the last fitting model (e.g., partial strong measurement invariance model vs. weak measurement invariance model). In the case of multiple parameters differing (e.g., item intercepts in more than one response condition at selection [t2]), it was also tested whether additional equality constraints between these parameters could be introduced (e.g., $\tau_{t2,h,1} = \tau_{t2,r,1}$). This was repeated until equivalence was achieved and the final models thereby specified. Within the final models, the equality of the latent trait retest correlations was then tested (i.e., $r_h = r_r = r_f$).

Results

Comparison across response conditions at selection (t1)

First, measurement invariance at selection (t1) was examined for the comparability of the three response conditions at selection (t1). The global fit statistics and model comparisons of the

Table 1. Model fits and model comparisons for the comparability of the three conditions at selection (t1), the measurement invariance (MI) across selection (t1) and retest (t2), and the comparison of latent retest correlations of the scale activity in familiar communicative situation.

			Model fit					Model comparison				
	No.	Model	χ^2	*df*	*p*	RMSEA	CFI	vs.	$\Delta\chi^2$	Δdf	*p*	ΔCFI
Comparability at selection (t1)	1a	Configural MI	85.213	87	.534	.000	1.000	—	—	—	—	—
	2a	Weak MI ($\lambda_{t1,h,1} = \lambda_{t1,r,1} = \lambda_{t1,f,1}$, etc.)	91.857	95	.572	.000	1.000	1a	6.659	8	.574	.000000
	3a	Strong MI ($\tau_{t1,h,1} = \tau_{t1,r,1} = \tau_{t1,f,1}$, etc.)	93.634	99	.633	.000	1.000	2a	2.616	4	.624	.000000
	4a	Strict MI ($\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t1,r,1} = \theta\varepsilon_{t1,f,1}$, etc.)	109.648	109	.465	.009	.998	3a	16.617	10	.083	.001997
	5a	4a + equal variances ($\Phi_{t1,h} = \Phi_{t1,r} = \Phi_{t1,f}$)	111.301	111	.474	.004	.999	4a	1.733	2	.421	.001435
	6a	5a + equal means ($\xi_{t1,h} = \xi_{t1,r} = \xi_{t1,f}$)	113.176	113	.478	.004	.999	5a	1.849	2	.397	.000019
MI across selection (t1) and retest (t2)	7a	Weak MI ($\lambda_{t1,h,1} = \lambda_{t2,h,1}$, etc.)	124.725	125	.490	.000	1.000	6a	11.807	12	.461	.000543
	8a	Strong MI ($\tau_{t1,h,1} = \tau_{t2,h,1}$, etc.)	155.095	140	.181	.036	.953	7a	46.733	15	.000	.046531
	9a	Partial strong MI	135.094	135	.482	.003	1.000	7a	10.720	10	.380	.000291
	10a	9a + equal residual variances ($\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t2,h,1}$, etc.)	144.528	150	.611	.000	1.000	9a	12.336	15	.653	.000291
	11a	10a + equal residual covariances ($rc_{1,h} = rc_{1,r} = rc_{1,f}$, etc.)	154.305	160	.612	.000	1.000	10a	9.825	10	.456	.000000
	12a	11a + equal variances ($\Phi_{t1,h} = \Phi_{t2,h}$, etc.)	171.275	163	.313	.025	.974	11a	8.770	3	.033	.025506
	13a	11a + partial equal variances	156.415	162	.609	.000	1.000	11a	1.970	2	.374	.000000
	14a	13a + equal means ($\xi_{t1,h} = \xi_{t2,h}$ etc.)	169.040	165	.398	.017	.988	13a	25.940	3	< .001	.012454
	15a	**13a + partial equal means**	**158.993**	**164**	**.596**	**.000**	**1.000**	**13a**	**3.048**	**2**	**.218**	**.000000**
	16a	15a + equal latent covariances ($r_h = r_r = r_f$)	170.974	166	.379	.019	.985	15a	5.992	2	.049	.015333

Note. The final model is shown in bold. RMSEA = root mean square error of approximation; CFI = comparative fit index.

successively more restricted models are shown in the upper parts of Table 1 (activity in familiar communicative situation), Table 2 (preventive health behavior in response to warning signals), and Table 3 (self-confidence in test situation). The results indicated that at selection (t1) strict measurement invariance across the three response conditions could be assumed for all scales (Model 1a—Model 4a, Model 1h—Model 4h, and Model 1s—Model 4s, respectively). Furthermore, adding the constraints for equal latent trait variances (Model 5a, Model 5h, and Model 5s, respectively) and equal latent trait means (Model 6a, Model 6h, and Model 6s, respectively) did not worsen the model fit. Therefore, the respondents randomly assigned to the three response conditions at retest (t2) did not differ in their response processes, latent trait mean, and latent trait variance at selection (t1) in all three scales.

Comparison across measurement points

Next, measurement invariance across selection (t1) and retest (t2) was tested to examine the extent to which faking behavior in the different response conditions affected the measurement properties of the scales. The global fit statistics of the successively more restricted models for testing measurement invariance across selection (t1) and retest (t2) and the respective model comparisons are summarized in the lower parts of Table 1 (activity in familiar communicative situation), Table 2 (preventive health behavior in response to warning signals), and Table 3 (self-confidence in test situation). The final parameter estimates are summarized in Table 4, and the latent retest correlations in Table 5.

Activity in familiar communicative situation

For the scale activity in familiar communicative situation (cf. Table 1), weak measurement invariance (Model 7a) could be assumed across selection (t1) and retest (t2). However, strong measurement invariance was not given (Model 8a): Four intercepts differed in the honest condition and one in the faking-good condition (Model 9a). The item intercepts differed complementarily: They were either higher in the honest condition at selection (t1) than at retest (t2), or lower in the faking-good condition at selection (t1) than at retest (t2). This violation of strong measurement invariance in the honest condition and faking-good condition was contrary to Hypotheses 1 and 4, respectively. Next, the equality of the residual variances was examined. This was the only scale where all residual variances were equal across the measurement points (Model 10a). Thus, strict measurement invariance was given in the reproduce condition, which supported Hypothesis 7. The residual covariances were also equal across the response conditions (Model 11a).

Next, the equality of the latent trait variances, means, and covariances was examined. The latent trait variances were not equal across the measurement points (Model 12a); the latent trait variance in the honest condition at selection (t2) was higher than all other latent trait variances (Model 13a). The latent trait means were also not equal (Model 14a), with the latent trait mean in the reproduce condition at selection (t2)

Table 2. Model fits and model comparisons for the comparability of the three conditions at selection (t1), the measurement invariance (MI) across selection (t1) and retest (t2), and the comparison of latent retest correlations of the scale preventive health behavior in response to warning signals.

			Model fit					Model comparison				
	No.	Model	χ^2	*df*	*p*	RMSEA	CFI	vs.	$\Delta\chi^2$	Δdf	*p*	ΔCFI
Comparability at selection (t1)	1h	Configural MI	77.624	87	.754	.000	1.000	—	—	—	—	—
	2h	Weak MI ($\lambda_{t1,h,1} = \lambda_{t1,r,1} = \lambda_{t1,f,1}$, etc.)	85.315	95	.752	.000	1.000	1h	7.666	8	.467	.000000
	3h	Strong MI ($\tau_{t1,h,1} = \tau_{t1,r,1} = \tau_{t1,f,1}$, etc.)	88.794	99	.759	.000	1.000	2h	3.510	4	.476	.000000
	4h	Strict MI ($\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t1,r,1} = \theta\varepsilon_{t1,f,1}$, etc.)	97.468	109	.778	.000	1.000	3h	8.716	10	.559	.000000
	5h	4h + equal variances ($\Phi_{t1,h} = \Phi_{t1,r} = \Phi_{t1,f}$)	98.092	111	.804	.000	1.000	4h	1.116	2	.572	.000000
	6h	5h + equal means ($\xi_{t1,h} = \xi_{t1,r} = \xi_{t1,f}$)	99.067	113	.822	.000	1.000	5h	0.707	2	.702	.000000
MI across selection (t1) and retest (t2)	7h	Weak MI ($\lambda_{t1,h,1} = \lambda_{t2,h,1}$, etc.)	120.555	125	.596	.000	1.000	6h	17.337	12	.137	.000000
	8h	Strong MI ($\tau_{t1,h,1} = \tau_{t2,h,1}$, etc.)	157.368	140	.150	.039	.948	7h	53.666	15	< .001	.052192
	9h	Partial strong MI	136.375	138	.523	.000	1.000	7h	17.772	13	.166	.000000
	10h	9h + equal residual variances ($\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t2,h,1}$, etc.)	178.211	153	.080	.045	.924	9h	37.509	15	.001	.075762
	11h	9h + partial equal residual variances	144.888	151	.625	.000	1.000	9h	10.269	13	.672	.000000
	12h	11h + equal residual covariances ($rc_{1,h} = rc_{1,r} = rc_{1,f}$, etc.)	158.733	161	.536	.000	1.000	11h	14.119	10	.168	.000000
	13h	12h + equal variances ($\Phi_{t1,h} = \Phi_{t2,h}$, etc.)	173.696	164	.287	.027	.971	12h	9.101	3	.028	.029137
	14h	12h + partial equal variances	158.230	163	.591	.000	1.000	12h	1.326	2	.515	.000000
	15h	14h + equal means ($\xi_{t1,h} = \xi_{t2,h}$ etc.)	228.389	166	.001	.068	.813	14h	94.058	3	< .001	.187485
	16h	**14h + partial equal means**	**160.633**	**164**	**.560**	**.000**	**1.000**	**14h**	**2.623**	**1**	**.105**	**.000000**
	17h	16h + equal latent covariances ($r_h = r_r = r_f$)	168.583	166	.430	.000	.992	16h	8.010	2	.018	.007763
	18h	16h + partial equal latent covariances	160.430	165	.586	.000	1.000	16h	0.001	1	.972	.000000

Note. The final model is shown in bold. RMSEA = root mean square error of approximation; CFI = comparative fit index.

Table 3. Model fits and model comparisons for the comparability of the three conditions at selection (t1), the measurement invariance (MI) across selection (t1) and retest (t2), and the comparison of latent retest correlations of the scale self-confidence in test situation.

			Model fit					Model comparison				
	No.	Model	χ^2	*df*	*p*	RMSEA	CFI	vs.	$\Delta\chi^2$	Δdf	*p*	ΔCFI
Comparability at selection (t1)	1s	Configural MI	89.598	87	.403	.019	.991	—	—	—	—	—
	2s	Weak MI ($\lambda_{t1,h,1} = \lambda_{t1,r,1} = \lambda_{t1,f,1}$, etc.)	96.012	95	.452	.011	.996	1s	6.442	8	.598	.005563
	3s	Strong MI ($\tau_{t1,h,1} = \tau_{t1,r,1} = \tau_{t1,f,1}$, etc.)	98.745	99	.488	.000	1.000	2s	3.241	4	.518	.003550
	4s	Strict MI ($\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t1,r,1} = \theta\varepsilon_{t1,f,1}$, etc.)	106.656	109	.546	.000	1.000	3s	7.634	10	.665	.000000
	5s	4s + equal variances ($\Phi_{t1,h} = \Phi_{t1,r} = \Phi_{t1,f}$)	109.475	111	.523	.000	1.000	4s	3.256	2	.196	.000000
	6s	5s + equal means ($\xi_{t1,h} = \xi_{t1,r} = \xi_{t1,f}$)	110.787	113	.541	.000	1.000	5s	1.210	2	.546	.000000
MI across selection (t1) and retest (t2)	7s	Weak MI ($\lambda_{t1,h,1} = \lambda_{t2,h,1}$, etc.)	137.937	125	.202	.036	.955	6s	21.971	12	.038	.045378
	8s	Partial weak MI	120.463	124	.573	.000	1.000	6s	9.817	11	.547	.000000
	9s	8s + equal intercepts ($\tau_{t1,h,1} = \tau_{t2,h,1}$, etc.)	148.558	139	.274	.029	.966	8s	36.997	15	.001	.033526
	10s	8s + partial equal intercepts	134.578	137	.543	.000	1.000	8s	15.309	13	.288	.000000
	11s	10s + equal residual variances ($\theta\varepsilon_{t1,h,1} = \theta\varepsilon_{t2,h,1}$, etc.)	335.831	152	.000	.122	.355	10s	278.787	28	< .001	.644831
	12s	10s + partial equal residual variances	144.253	147	.549	.000	1.000	10s	24.177	23	.394	.000000
	13s	12s + equal residual covariances ($rc_{1,h} = rc_{1,r} = rc_{1,f}$, etc.)	152.073	157	.596	.000	1.000	12s	7.622	10	.666	.000000
	14s	13s + equal variances ($\Phi_{t1,h} = \Phi_{t2,h}$, etc.)	193.861	160	.035	.051	.881	13s	79.337	3	< .001	.118776
	15s	**13s + partial equal variances**	**153.800**	**158**	**.580**	**.000**	**1.000**	**13s**	**1.909**	**1**	**.167**	**.000000**
	16s	15s + equal means ($\xi_{t1,h} = \xi_{t2,h}$ etc.)	284.375	161	< .001	.097	.567	15s	158.876	3	< .001	.432769
	17s	15s + equal latent covariances ($r_h = r_r = r_f$)	167.275	160	.331	.024	.974	15s	13.207	2	.001	.025520
	18s	15s + partial equal latent covariances	154.015	159	.597	.000	1.000	15s	0.280	1	.597	.000000

Note. The final model is shown in bold. RMSEA = root mean square error of approximation; CFI = comparative fit index.

Table 4. Comparison of selected unstandardized parameter estimates of selection (t1) to the three conditions at retest (t2) for all three scales. For selection (t1), the unstandardized parameters are given. For re-text (t2), either the unstandardized parameters are given, or = t1 in case of invariance across measurement points.

	t1 all conditions	t2 honest	t2 reproduce	t2 faking-good		t1 all conditions	t2 honest	t2 reproduce	t2 faking-good
Activity in familiar communicative situation									
i1-intercept ($\tau_{tx,c,1}$)	3.872	3.718	= t1	= t1	i1-res.-var. ($\theta\varepsilon_{tx,c,1}$)	0.090	= t1	= t1	= t1
i2-intercept ($\tau_{tx,c,2}$)	3.368	3.167	= t1	= t1	i2-res.-var. ($\theta\varepsilon_{tx,c,2}$)	0.288	= t1	= t1	= t1
i3-intercept ($\tau_{tx,c,3}$)	3.746	3.560	= t1	= t1	i3-res.-var. ($\theta\varepsilon_{tx,c,3}$)	0.160	= t1	= t1	= t1
i4-intercept ($\tau_{tx,c,4}$)	3.031	= t1	= t1	3.336	i4-res.-var. ($\theta\varepsilon_{tx,c,4}$)	0.370	= t1	= t1	= t1
i5-intercept ($\tau_{tx,c,5}$)	3.629	3.443	= t1	= t1	i5-res.-var. ($\theta\varepsilon_{tx,c,5}$)	0.236	= t1	= t1	= t1
Latent trait mean ($\xi_{tx,c}$)	−0.009	= t1	−0.098	= t1	Latent trait variance ($\Phi_{tx,c}$)	0.060	0.145	= t1	= t1
Preventive health behavior in response to warning signals									
i1-intercept ($\tau_{tx,c,1}$)	3.232	= t1	= t1	= t1	i1-res.-var. ($\theta\varepsilon_{tx,c,1}$)	0.334	= t1	= t1	= t1
i2-intercept ($\tau_{tx,c,2}$)	3.577	= t1	= t1	= t1	i2-res.-var. ($\theta\varepsilon_{tx,c,2}$)	0.232	= t1	= t1	= t1
i3-intercept ($\tau_{tx,c,3}$)	3.765	= t1	= t1	= t1	i3-res.-var. ($\theta\varepsilon_{tx,c,3}$)	0.104	0.236	0.236	= t1
i4-intercept ($\tau_{tx,c,4}$)	3.686	= t1	3.433	= t1	i4-res.-var. ($\theta\varepsilon_{tx,c,4}$)	0.238	0.386	= t1	= t1
i5-intercept ($\tau_{tx,c,5}$)	3.454	= t1	= t1	3.615	i5-res.-var. ($\theta\varepsilon_{tx,c,5}$)	0.208	= t1	= t1	= t1
Latent trait mean ($\xi_{tx,c}$)	0.005	−0.294	−0.294	0.161	Latent trait variance ($\Phi_{tx,c}$)	0.125	= t1	0.280	= t1
Self-confidence in test situation									
i1-intercept ($\tau_{tx,c,1}$)	3.989	= t1	= t1	= t1	i1-res.-var. ($\theta\varepsilon_{tx,c,1}$)	0.176	= t1	= t1	0.048
i2-intercept ($\tau_{tx,c,2}$)	3.959	= t1	= t1	= t1	i2-res.-var. ($\theta\varepsilon_{tx,c,2}$)	0.238	= t1	= t1	0.077
i3-intercept ($\tau_{tx,c,3}$)	3.725	= t1	= t1	3.881	i3-res.-var. ($\theta\varepsilon_{tx,c,3}$)	0.280	= t1	= t1	0.137
i4-intercept ($\tau_{tx,c,4}$)	3.936	= t1	= t1	= t1	i4-res.-var. ($\theta\varepsilon_{tx,c,4}$)	0.272	= t1	= t1	0.060
i5-intercept ($\tau_{tx,c,5}$)	4.004	3.837	3.837	= t1	i5-res.-var. ($\theta\varepsilon_{tx,c,5}$)	0.222	= t1	= t1	0.007
Latent trait mean ($\xi_{tx,c}$)	−0.569	−0.820	−0.685	−0.060	Latent trait variance ($\Phi_{tx,c}$)	0.193	0.403	= t1	0.069

Note. The subscript *tx* denotes t1 or t2, and the subscript *c* the response conditions honest, reproduce, or faking-good. res.-var. = residual variance.

being lower than all other latent trait means (Model 15a). The latent trait mean difference in the reproduce condition was contrary to Hypothesis 9, and the equal latent trait means in the honest condition and faking-good condition were contrary to Hypotheses 3 and 6, respectively. Finally, the latent retest correlations (cf. Table 5) differed across the three response conditions (Model 16a): The latent trait retest correlation was highest in the reproduce condition (supporting Hypothesis 8), lower in the honest condition (supporting Hypothesis 2), and negligible in the faking-good condition (supporting Hypothesis 5). This was the only scale where no equality of the latent retest correlations could be established (cf. Table 5).

Preventive health behavior in response to warning signals

For the scale preventive health behavior in response to warning signals (cf. Table 2), weak measurement invariance was given across the measurement points (Model 7h), whereas strong measurement invariance was not (Model 8h). Two intercepts differed (Model 9h), one in the faking-good condition (t1 < t2; not supporting Hypothesis 4), and one in the reproduce condition (t1 > t2; not supporting Hypothesis 7). Next, equality constraints on the residual variances were imposed (Model 10h), which showed that the residual variances differed across the measurement points. Overall, three residual variances were lower at selection (t1) than at retest (t2); the residual variances of one item in the honest condition and reproduce condition to the same extent, and additionally another residual variance in the honest condition (Model 11h). These differences in the residual variances in the honest condition were contrary to Hypothesis 1. The covariances of the residuals were equal across all three response conditions (Model 12h).

Afterward, equality constraints for the latent trait variances, means, and covariances were imposed. The latent trait variances were not equal across the measurement points (Model 13h), with the variance of the latent trait in the reproduce condition at retest (t2) being the highest (Model 14h). The latent trait means also differed across the measurement points (Model 15h): In the honest condition and the reproduce condition they were equally higher at selection (t1) than at retest (t2), and lower in the faking-good condition (Model 16h). The latent trait mean differences in the honest condition and faking-good condition offered support for Hypotheses 3 and 6, respectively. However, the latent trait mean difference in the reproduce condition was contrary to Hypothesis 9. Finally, the latent retest correlations were also not equal across all three conditions (Model 17h), but were across the honest condition and faking-good condition (Model 18h). As can be seen in Table 5, the latent retest correlations were moderate in the honest condition and faking-good condition (supporting Hypothesis 2 and not supporting Hypothesis 5, respectively), and higher in the reproduce condition (supporting Hypothesis 8).

Table 5. The latent retest correlations of the final models of all three scales.

	Condition		
Scale	Honest	Reproduce	Faking-good
Activity in familiar communicative situation	.497	.898	.019
Preventive health behavior in response to warning signals	.387*	.699	.395*
Self-confidence in test situation	.533*	.442*	.121

Note. Per scale, the not differing latent retest correlations are marked (*).

Self-confidence in test situation

The scale self-confidence in test situation (cf. Table 3) differed from the other two scales, as weak measurement invariance was not given (Model 7s). Subsequent analyses indicated that the factor loading of one item of this scale was higher in the faking-good condition than in all other response conditions

(Model 8s). This violation of weak measurement invariance in the faking-good condition contradicted Hypothesis 4. Similar to the other two scales, constraining the item intercepts to be equal across the measurement points decreased the model fit (Model 9s). One intercept differed per response condition (Model 10s): One intercept was equally higher in the honest condition and reproduce condition at selection (t1) than at retest (t2), whereas in the faking-good condition one item intercept was lower at selection (t1) than at retest (t2). The differences in intercepts in the honest condition and reproduce condition were contrary to Hypotheses 1 and 7, respectively. The residual variances were also not equal across measurement points (Model 11s). No residual variances were affected in the honest condition nor the reproduce condition, but all residual variances in the faking-good condition were higher at selection (t1) than at retest (t2). Regarding the residual covariances, they were equal across all three response conditions (Model 13s).

Finally, the equality of the latent trait variances, means, and covariances was examined. As with the other two scales, the latent trait variances were not equal across the measurement points (Model 14s). The latent trait variance was lower at selection (t1) than at retest (t2) in the honest condition, whereas it was higher in the faking-good condition (Model 15s). The latent trait means also differed (Model 16s): The latent trait means in the honest condition and reproduce condition were lower at selection (t1) than at retest (t2), whereas it was higher at selection (t1) than at retest (t2) in the faking-good condition. Similar to the results of the scale preventive health behavior in response to warning signals, these latent trait mean differences supported Hypotheses 3 and 6, although contrary to Hypothesis 9. This was the only scale where no further equality of the latent trait means could be established. The latent retest correlations were also unequal (Model 17s), but were equal across the honest condition and reproduce condition (Model 18s). The latent retest correlations (cf. Table 5) were moderate in the honest condition and reproduce condition (supporting Hypothesis 2, and not supporting Hypothesis 8 respectively) and lower in the faking-good condition (supporting Hypothesis 5).

Discussion

Although it has been examined whether, and to what extent, applicants can fake personality scales, more research is needed to evaluate the effect of different kinds of faking behavior on the response process involved (Goffin & Boyd, 2009; Griffith & Peterson, 2011; Kuncel et al., 2011). This study is expected to shed some light on applicants' faking behavior by retesting applicants as incumbents in three response conditions: (a) honest responses, (b) faked-good responses, and (c) responses given when attempting to reproduce one's own responses from when applying. Measurement invariance, retest correlations, and latent trait mean differences across these three response conditions and applicants' response behavior in a real-life selection setting were examined.

Honest, applicants', and instructed faking responses

It was hypothesized that the response behavior of applicants is comparable to the response behavior of incumbents completing the personality scale in an honest condition (cf. Robie et al., 2001). The findings of this study only partially support this hypothesis. None of the scales exhibited strict measurement invariance across applicants in a real-life selection setting and incumbents responding honestly (contradictory to Hypothesis 1). Faking in a real-life selection setting caused respondents primarily to increase their responses to selected items. As this was not the case for all items to the same extent, strong measurement invariance was violated for two out of three scales. For the one scale where strong measurement invariance was given, strict measurement invariance was nevertheless violated. This decrease of unaccounted for variance from honest response behavior to applicants' response behavior seems to be attributable to real-life selection settings reducing the potential ambiguity of item contents. If an item's content is ambiguous, this would introduce variance unaccounted for by the psychometric model. However, being in a real-life selection setting could reduce this ambiguity, because respondents will less likely relate the item content to all possible settings, but rather to the job-relevant settings.

Nevertheless, our results corroborate that the detrimental effects of applicants' faking behavior should not be overestimated (e.g., Bradley & Hauenstein, 2006; Hogan et al., 2007; Ones et al., 2007), as some measurement properties were consistently preserved (configural and weak measurement invariance). The preservation of weak measurement invariance also indicates that predictive validity coefficients might not be detrimentally affected (Chen, 2008). This might explain why the effects of applicants' faking behavior has been deemed to be negligible in real-life selection settings based on a comparison of predictive validity coefficients across honest responding conditions and real-life selection settings (for an overview, see Ones et al., 2007).

The finding of a lack of strict measurement invariance across applicants' response behavior in a real-life selection setting and the honest response behavior contradicts our hypothesis and the results of Robie et al. (2001) on which this hypothesis was based. However, there was one main difference between this study and the study conducted by Robie et al. In the selection situation of Robie et al. (2001), a warning was presented "that distorted self-descriptions would invalidate the … test results" (p. 195). No such warning was presented in this study. Prior research indicated that applicants wrongly fear their faking behavior might be exposed (König et al., 2012). Furthermore, it has been shown that a warning that faking can be detected reduces faking when the negative consequences of being detected are emphasized (e.g., Dwight & Donovan, 2003; Goffin & Woods, 1995). Therefore, applicants might have been more reluctant to fake in the study conducted by Robie et al. (2001) than in our study.

In summary, the effects of applicants' faking behavior were not negligible in our study. Applicants' faking behavior compromised either strong or strict measurement invariance for all three scales. Furthermore, applicants' faking behavior distorted the rank-orders (Hypothesis 2) to a modest extent ($r = .497$, $r = .387$, and $r = .533$). This distortion was in line with previous research (e.g., Griffin & Wilson, 2012; Griffith et al., 2007; Peterson et al., 2011) and our hypotheses (Hypothesis 2), and corroborates the effect of detrimental applicants' faking behavior on selection decisions.

The response behavior in an instructed faking-good condition was hypothesized to be different from other response behaviors, such as applicants' response behavior and honest response behavior (e.g., Miller & Ruggs, 2014; Robie et al., 2001). The findings of this study corroborate this hypothesis. This difference could be seen in the distortion of rank-order (Hypothesis 5). This distortion was in line with previous studies (e.g., Ellingson et al., 1999; Griffith et al., 2007), and showed that the rank-order could be completely distorted ($r = .019$), but was at its highest still only modestly preserved ($r = .395$). Furthermore, instructed faking compared to applicants' response behavior consistently violated strong measurement invariance (Hypothesis 4). This was always the case for items that were not yet affected by applicants' faking behavior, underlining that applicants engage in less faking than they could.

For one scale, weak and strict measurement invariance were additionally violated across applicants' response behavior and instructed faking. This was the scale (self-confidence in test situation) for which instructed faking led to a ceiling effect. Due to the limited variance in the items of this scale, the unaccounted variance of all items was also decreased. For the item with the most pronounced restriction of variance, even the factor loading was affected. This was also the item that had the highest average responses at selection (t1). The item content of this item was related to being stable, a trait for which even an overly high endorsement cannot be viewed negatively for teachers.

Taken together, our findings suggest that instructed faking will always detrimentally affect the measurement properties of scales. Moreover, this detrimental effect will be even more pronounced when instructed faking leads to strong ceiling effects. Consequently, caution is advised when making selection decisions: When faking behavior causes a ceiling effect—for example, applicants are highly motivated to fake by very high selection ratios—the selection decisions could be heavily biased.

We hypothesized that we would find the lowest latent means for honest response behavior, higher latent means for applicants' response behavior, and even higher latent means for instructed faking-good response behavior (cf. Birkeland et al, 2006; Viswesvaran & Ones, 1999). However, these hypotheses were only partially confirmed (Hypotheses 3 & 6). The latent means of two of the three scales showed this expected pattern, whereas the latent means of the remaining scale (activity in familiar communicative situation) did not differ significantly across response conditions. Because of this, we did a post-hoc comparsion of this remaining scale's manifest raw means. This post-hoc comparsion revealed the honest response behavior having a lower raw mean than the applicants' response behavior ($d = 0.36$), which in turn had a lower raw mean than the instructed faking-good response behavior ($d = 0.33$). Like the aforementioned meta-analyses (Birkeland et al, 2006; Viswesvaran & Ones, 1999), most studies on mean differences due to faking behavior only examine the mean difference of raw scores. Unfortunately, our finding that strong measurement invariance was not consistently given indicated that such mean raw score comparisons might be biased (Li & Zumbo, 2009). If strong measurement invariance is unaffected or affected only partially, however, the latent trait means will nevertheless show that mean honest responses are lower than mean applicants' responses, which are lower than mean instructed faking responses. Therefore, raw mean score differences due to faking behavior might reflect true differences, but, in the most detrimental case, might only reflect a uniform measurement bias.

Stability of applicants' faking behavior

To try to shed even more light on applicants' faking behavior, this study also addressed the stability of applicants' faking behavior over half a year. Based on previous findings, we hypothesized applicants' faking behavior to be relatively stable (cf. Hogan et al., 2007). Our results demonstrated that half a year later, incumbents could partially reproduce their response process from when they were applying (partial support for Hypothesis 7). One scale exhibited strict measurement invariance (activity in familiar communicative situation), but for the other two scales, only partial strong measurement invariance was given. Additionally, an increase in unaccounted for variance by the psychometric model was observed. For these affected items, the effects were consistent with an honest response behavior. Incumbents also consistently underestimated their faking intensity from when applying; that is, their latent means decreased (contradictory to Hypothesis 9). By contrast, the rank-orders (Hypothesis 8) could remain remarkably stable ($r = .90$ and $r = .699$), but were at least equally high as when compared to honest response behavior ($r = .442$). Consequently, studies using test scores obtained from incumbents who were instructed to respond like applicants might underestimate the effects of applicants' faking behavior (e.g., Fell & König, 2016).

In summary, our results suggest that incumbents seem to be able to reproduce their response behavior from a real-life selection setting, and when in doubt, they rely on an honest responding behavior as a frame of reference. Given that half a year had passed between the real-life selection setting and the attempt to reproduce one's own faking behavior, it is remarkable that only a few measurement properties could not be reproduced. Therefore, these findings suggest that the individual differences determinants of faking behavior might be predominantly dispositional, as was already postulated by Snell et al. (1999). However, respondents underestimated their faking intensity when asked to reproduce their faking behavior from a real-life selection setting. In consequence, situational factors might also play a nonnegligible role.

Limitations and conclusion

Although this study aimed to provide insights into applicants' response behavior, it is, of course, not without limitations. First, only one personality inventory was used in this study. Future research might consider using more than one personality inventory to examine the effects of faking in cases where scales differ in structure, response format, and their perceived relevance for the particular job. The perceived relevance of scales for a particular job might also be directly assessed in future studies, as the appraisal of relevance could differ between applicants and scholars. Also, more traditional Big Five inventories could be used, which do not relate the items to a situational

context. However, it should be noted that the personality inventory used in this study behaved consistently with faking effects shown for more widely used personality inventories.

Future studies might also wish to experimentally manipulate the selection ratio to examine the effects of the selection ratio on the propensity and intensity to fake. However, differences in selection ratios could affect the extent of range restrictions due to the selection process (Linn, 1968). This might lead to an underestimation of the correlations (Bortz & Döring, 2006). To avoid such selection effects, future studies might also want to include rejected applicants in their samples. We do not believe that a range restriction significantly biased our results, as all three response conditions should be equally affected by this problem.

It would also be favorable to conduct future studies with larger sample sizes. The sample size in this study was rather small per group for measurement invariance analysis. Due to this concern, subsequent Monte Carlo simulation studies were conducted to examine if our parameter estimates were biased. Based on the final models, the Monte Carlo simulation studies suggested that the parameter estimates were largely unbiased, and that their standard errors were only slightly biased for two items of the scale activity in familiar communicative situation. A detailed summary of the Monte Carlo simulation studies is available from the first author on request.

Furthermore, it remains unclear what caused respondents' underestimation of their own faking intensity when attempting to reproduce their applicants' response behavior. Future research should examine whether there are differences in the selection settings and individual differences in the applicants by which the faking intensity can be reproduced with more or with less accuracy. Finally, it is always possible that responses in an honest condition might still not be entirely honest.

Despite these limitations, this study improved our understanding of applicants' response behavior in real-life situations. This study indicates that applicants' response behavior in a real-life situation is quite stable and faking primarily affects strong measurement invariance by inducing a uniform bias as compared to an honest responding behavior, which in turn detrimentally effects the stability of the rank-order of the respondents according to their standing on the latent trait. However, it should be noted that these effects are by far less severe than the effects of an instructed faking.

References

Arendasy, M., Sommer, M., & Feldhammer, M. (2011). *Manual: Big Five Structure Inventory* (BFSI). Mödling, Austria: Schuhfried.

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias: A simulation study. *AStA Advances in Statistical Analysis*, *94*, 117–127.

Birkeland, S. A., Manson, T. M., Kisamore, J., 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*, 317–356.

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluationfür Human- und Sozialwissenschaftler* [Research methods and evaluation for the social sciences] (4th rev. ed.). Heidelberg, Germany: Springer.

Bradley, K. M., & Hauenstein, N. M. (2006). The moderating effects of sample type as evidence of the effects of faking on personality scale correlations and factor structure. *Psychology Science*, *48*, 313–335.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*, 1005–1018.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness of fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Costa, P. T., & McCrae, R. R. (1992). *Neo PI–R professional manual*. Odessa FL: Psychological Assessment Resources.

Curran, R. J., West, S. G., & Finch, J. E. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*, 16–29.

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*, 209–225.

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

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, *16*, 1–23.

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

Fell, C. B., & König, C. J. (2016). Cross-cultural differences in applicant faking on personality tests: A 43-nation study. *Applied Psychology*, *65*, 671–717.

Ferrando, P. J., & Anguiano-Carrasco, C. (2009a). Assessing the impact of faking on binary personality measures: An IRT-based multiple-group factor analytic procedure. *Multivariate Behavioral Research*, *44*, 497–524.

Ferrando, P. J., & Anguiano-Carrasco, C. (2009b). The interpretation of the EPQ Lie scale scores under honest and faking instructions: A multiple-group IRT-based analysis. *Personality and Individual Differences*, *46*, 552–556.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology*, *50*, 151–160.

Goffin, R. D., & Woods, D. M. (1995). Using personality testing for personnel selection: Faking and test−taking inductions. *International Journal of Selection and Assessment*, *3*, 227–236.

Gold, Y., & Roth, R. A. (1993). *Teachers managing stress and preventing burnout: The professional health solution*. London, UK: Routledge.

Griffin, B., & Wilson, I. G. (2012). Faking good: Self-enhancement in medical school applications. *Medical Education*, *46*, 485–490.

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

Griffith, R. L., & Peterson, M. H. (2011). One piece at a time: The puzzle of applicant faking and a call for theory. *Human Performance*, *24*, 291–301.

Hartman, N. S., & Grubb, W. L. (2011). Deliberate faking on personality and emotional intelligence measures. *Psychological Reports*, *108*, 120–138.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, *92*, 1270–1285.

Hogan, R. (1992). *Hogan Personality Inventory manual* (3rd ed.). Tulsa, OK: Hogan Assessment Systems.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology*, *93*, 140–154.

König, C. J., Merz, A.-S., & Trauffer, N. (2012). What is in applicants' minds when they fill out a personality test? Insights from a qualitative study. *International Journal of Selection and Assessment*, *20*, 442–452.

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal*, *6*, 151–162.

Koschmieder, C., Pretsch, J., & Neubauer, A. (2015, July). *Emotional intelligence, personality and general mental ability in teacher student selection: An examination of predictive validity and overlap*. Paper presented at the Conference of the International Society for the Study of Individual Differences (ISSID) in London, ON, Canada.

Krammer, G., & Pflanzl, B. (2015). Faking von Persönlichkeitseigenschaften bei Zulassungsverfahren für Lehramtsstudien. [Faking of personality measures for college admission in teacher education.] *Zeitschrift für Pädagogische Psychologie*, *29*, 205–214.

Krammer, G., Sommer, M., & Arendasy, M. E. (2016). Realistic job expectations predict academic achievement. *Learning and Individual Differences*, *51*, 341–348.

Kuncel, N. R., Goldberg, L. R., & Kiger, T. (2011). A plea for process in personality prevarication. *Human Performance*, *24*, 373–378.

Kuncel, N. R., & Tellegen, A. (2009). A conceptual and empirical reexamination of the measurement of the social desirability of items: Implications for detecting desirable response style and scale development. *Personnel Psychology*, *62*, 201–228.

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions. *Psicológica*, *30*, 343–370.

Linn, R. L. (1968). Range restriction problems in the use of self-selected groups for test validiation. *Psychological Bulletin*, *69*, 69–73.

Marcus, B. (2009). "Faking" from the applicant's perspective: A theory of self-presentation in personnel selection settings. *International Journal of Selection and Assessment*, *17*, 417–430.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cut off values for fit indexes and dangers in overgeneralizing Hu & Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

Mayr, J. (2011). Der Persönlichkeitsansatz der Lehrerforschung. [The personality approach to research in the field of teaching.] In E. Terhart, H. Bennewitz & M. Rothland (Eds.), *Handbuch der Forschung zum Lehrerberuf* [A research guide to the teaching profession] (pp. 125–148). Münster, Germany: Waxmann.

Mayr, J., & Brandstätter, H. (1998). *Lehrer/in werden?* [Becoming a teacher?] Wien, Germany: Bundesministerium für Unterricht und kulturelle Angelegenheiten.

McCrae, R. R., Yik, M. S., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology*, *74*, 1041.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*, 812–821.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology*, *36*, 979–1016.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, *93*, 568–592.

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

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.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Mislevy, R. J., Haertel, G., Cheng, B. H., Ructtinger, L., DeBarger, A., Murray, E., … Vendlinski, T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*, 121–140.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C., III. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, *88*, 348–355.

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

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*, 995–1027.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (pp. 125–167). Amsterdam, The Netherlands: Elsevier.

Peterson, M. H., Griffith, R. L., Isaacson, J. A., O'Connell, M. S., & Mangos, P. M. (2011). Applicant faking, social desirability, and the prediction of counterproductive work behaviors. *Human Performance*, *24*, 270–290.

Raykov, T., Marcoulides, G. A., & Li, C. H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, *72*, 954–974.

R Core Team. (2015). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria: R Core Development Team. Retrieved from http://www.R-project.org/

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance*, *14*, 187–207.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, *83*, 634–644.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Ryan, A. M., & Boyce, A. S. (2006). What do we know and where do we go? Practical directions for faking research. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 357–371). Greenwich, CT: Information Age.

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 167–180.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Schaarschmidt, U., & Fischer, A. (2013). *Manual Inventar zur Persönlichkeitsdiagnostik in Situationen* [Inventory for personality assessment in situations] (Version 21—Revision 2). Mödling, Austria: SCHUHFRIED GmbH.

Schulz-Kolland, R., Krammer, G., Rottensteiner, E., & Weitlaner, R. (2014). Die Validität von Zulassungsverfahren—Befunde der Pädagogischen Hochschule Steiermark. [On the validity of college admission processes—findings of the University College of Teacher Education Styria.] *Neuen @Hochschul-Zeitung*, *3*, 85–88.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review*, *9*, 219–242.

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

Unterbrink, T., Hack, A., Pfeifer, R., Buhl-Grießhaber, V., Müller, U., Wesche, H., … Bauer, J. (2007). Burnout and effort–reward imbalance in a sample of 949 German teachers. *International Archives of Occupational and Environmental Health*, *80*, 433–441.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.

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

Winkelspecht, C., Lewis, P., & Thomas, A. (2006). Potential effects of faking on the NEO PI–R: Willingness and ability to fake changes who gets hired in simulated selection decisions. *Journal of Business and Psychology*, *21*, 243–259.

Zickar, M. J., & Robie, C. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, *84*, 551–563.

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

Ziegler, M., & Bühner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, *69*, 548–565.