Schulte, Niklas; Kaup, Lucas; Bürkner, Paul-Christian; Holling, Heinz

The Fakeability of Personality Measurement with Graded Paired Comparisons

Journal of Business and Psychology

Provided in Cooperation with:
Springer Nature

Suggested Citation: Schulte, Niklas; Kaup, Lucas; Bürkner, Paul-Christian; Holling, Heinz (2024) : The Fakeability of Personality Measurement with Graded Paired Comparisons, Journal of Business and Psychology, ISSN 1573-353X, Springer US, New York, NY, Vol. 39, Iss. 5, pp. 1067-1084, https://doi.org/10.1007/s10869-024-09931-0

Journal of Business and Psychology (2024) 39:1067–1084
https://doi.org/10.1007/s10869-024-09931-0

ORIGINAL PAPER

The Fakeability of Personality Measurement with Graded Paired Comparisons

Niklas Schulte[1] · **Lucas Kaup**[2] · **Paul-Christian Bürkner**[3] · **Heinz Holling**[2]

Accepted: 1 January 2024 / Published online: 7 February 2024

Abstract

This study compares the faking resistance of Likert scales and graded paired comparisons (GPCs) analyzed with Thurstonian IRT models. We analyzed whether GPCs are more resistant to faking than Likert scales by exhibiting lower score inflation and better recovery of applicants' true (i.e., honest) trait scores. A total of $N = 573$ participants completed either the Likert or GPC version of a personality questionnaire first honestly and then in an applicant scenario. Results show that participants were able to increase their scores in both the Likert and GPC format, though their score inflation was smaller in the GPC than the Likert format. However, GPCs did not exhibit higher honest–faking correlations than Likert scales; under certain conditions, we even observed negative associations. These results challenge mean score inflation as the dominant paradigm for judging the utility of forced-choice questionnaires in high-stakes situations. Even if forced-choice factor scores are less inflated, their ability to recover true trait standings in high-stakes situations might be lower compared with Likert scales. Moreover, in the GPC format, faking effects correlated almost perfectly with the social desirability differences of the corresponding statements, highlighting the importance of matching statements equal in social desirability when constructing forced-choice questionnaires.

Keywords Forced-choice · Thurstonian IRT model · Ipsative data · Graded paired comparisons · Graded-preference items · Compositional items IRT

Forced-choice questionnaires have been used for a long time to prevent faking and other response distortions (Cao & Drasgow, 2019; Jackson et al., 2000; Saville & Willson, 1991). However, recent studies show that discrete forced-choice formats often yield reliabilities that are too low for individual diagnostics and, moreover, allow comparisons between individuals only to a limited extent (Bürkner et al., 2019; Schulte et al., 2020). One reason for this is that psychometrically, the selection or rejection of a forced-choice item within an item block provides relatively less item information regarding the trait score, compared to rating (i.e., Likert) items.

A promising solution that might combine the high information generated by Likert items with the faking resistance of the forced-choice format are graded paired comparisons (GPCs). In a GPC, two items are placed on both ends of a rating scale and respondents have to indicate the degree to which they prefer one item over the other (De Beuckelaer et al., 2013; Huber & Holbrook, 1982). Thus, they are a combination of Likert-type rating items and conventional dichotomous forced-choice tasks with two items. The two included items maintain the characteristic property of forced-choice scales in that not all items can be fully endorsed simultaneously, because one item's gain is the alternative item's loss. At the same time, the graded response provides information about the relative preference of one item over another, increasing the information about the latent traits. Figure 1 shows a GPC sample item.

So far, little is known about GPCs as a response format for personality measures. This pre-registered study was designed to investigate the potential merits of GPCs over

Additional supplementary materials may be found here https://osf.io/fx4yz/.

✉ Niklas Schulte
n.schulte@fu-berlin.de

[1] Department of Education and Psychology, Freie Universität Berlin, 14195 Berlin, Germany

[2] University of Münster, Münster, Germany

[3] TU Dortmund University, Dortmund, Germany

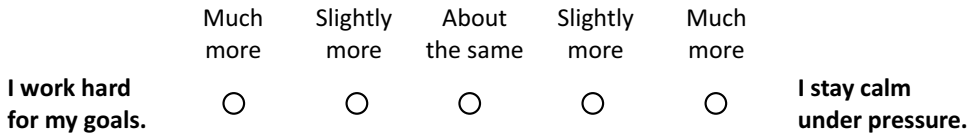


Fig. 1 Example of a graded paired comparison (GPC)

Likert items regarding their ability to resist faking attempts in the context of personnel selection. Additionally, it investigates whether the expected reliability gains are sufficient to achieve a measurement error small enough for individual diagnostic purposes in practice. Although a recent meta-analysis emphasizes the influence of item desirability on the fakeability of forced-choice questionnaires (Cao & Drasgow, 2019), little is known about what to consider when trying to match items on desirability during the construction of forced-choice item blocks. To inform test constructors about desirability matching, as it is perhaps the most important feature of forced-choice tests, we will investigate in detail the effects of item and trait desirability as well as the association with the keyed direction of items (i.e., the sign of the factor loading).

Forced-Choice Questionnaires as a Potential Remedy for the Problems Associated with Self-Reports

Questionnaires designed to measure personality, attitudes, interests, or values usually use self-report measures with Likert scales. On a Likert scale (Likert, 1932), test-takers have to rate their agreement to self-describing statements, i.e., from 1 (*strongly agree*) to 5 (*strongly disagree*). However, these items are subject to a number of biases, such as faking in a socially desirable direction, acquiescence (confirmation tendency), exaggerated coherence between items of different traits ("halo" effect), and several others (Paulhus & Vazire, 2007; Wetzel & Greiff, 2018; Wetzel et al., 2016).

These distortions may compromise the validity of the conclusions derived from Likert scales (Christiansen et al., 2005), especially in high-stakes situations, such as personnel selection (Birkeland et al., 2006; Christiansen et al., 2005) and performance appraisal (Brown et al., 2017), clinically relevant constructs (Young, 2018), or when the measured traits are particularly undesirable, such as "dark" personality traits (Guenole et al., 2018; Paulhus & Jones, 2014). Response biases can also be problematic when differentiation is particularly important, such as in market research or career advice (Wang et al., 2017). Problems can also arise in applications like comparative cultural research, where the groups under comparison may differ in these biases (J. Lee et al., 2002).

As a solution to these issues, forced-choice response formats have been suggested, in which respondents have to decide between two or more items, thus preventing many Likert scale biases by design (e.g., Hontangas et al., 2015; Jackson et al., 2000; Saville & Willson, 1991; Wetzel et al., 2016). To distinguish between forced-choice format types, we use the term discrete forced-choice format to refer to the classical one that can be transferred into binary comparisons (but can consist of more than two items) and the term GPC for two items with a rating scale in between. Discrete forced-choice items, however, are associated with two central problems: ipsativity and loss of information in comparison with Likert scales. A person parameter estimate is ipsative if all measured dimensions add up to the same total for each individual (Clemans, 1966). Therefore, a respondent's score on one dimension depends on their score on all other dimensions, making inter-individual comparisons based on ipsative scores questionable (Cattell, 1944; Hicks, 1970). This is highly problematic for many applications of forced-choice questionnaires, e.g., in personnel selection, where the comparison between applicants is the central objective (Johnson et al., 1988). Quasi-ipsative measures (e.g., Rasheed & Robie, 2023) only partially solve these problems.

To derive normative parameter estimates from forced-choice responses, Thurstonian IRT models were introduced (Brown & Maydeu-Olivares, 2011). An estimate is normative if it indicates the relative trait level compared to other subjects of the population distribution (Cattell, 1944). Nevertheless, the loss of information relative to Likert items remains problematic. Simulation studies suggest that even when scored with Thurstonian IRT models, conventional forced-choice items will yield low reliabilities in most applied conditions (Bürkner et al., 2019; Schulte et al., 2020). Because in GPCs respondents indicate their preference for one item over the other on a rating scale instead of only discretely selecting one of them, more information is captured about their expression on latent traits. This is a promising starting point to achieve higher reliabilities compared to discrete forced-choice methods.

Graded Paired Comparisons as a Psychometric Method

In the past, GPCs have primarily been used to investigate consumer preferences (Agresti, 1992; Alfaro-Rodriguez et al., 2005; De Beuckelaer et al., 2013; Ofir, 2004). For example, in marketing research, GPCs are often employed in the context of conjoint analyses (e.g., Scholz et al., 2010). GPCs are also known by slightly different names, such as

constant sum paired comparisons (Skedgel et al., 2015) or ordinal paired comparisons (Agresti, 1992).

Recently, Brown and Maydeu-Olivares (2018) proposed applying GPCs in personality assessments, as GPCs entail advantages over discrete forced-choice formats: First, while discrete forced-choice formats deteriorate participant reactions (Borman et al., 2023), GPCs can lead to greater acceptance by participants, as participants are not as harshly forced to choose between options that might describe them equally (in)appropriately (Dalal et al., 2019). Second, the aforementioned information gain achieved by re-introducing a rating scale might help to increase reliability. Indeed, simulations demonstrate higher levels of reliability for graded compared with discrete forced-choice formats (Lingel et al., 2022). Unfortunately, re-implementing a Likert scale also brings back typical method-related response biases such as extreme responding, but empirical results suggest these effects to be negligible (De Beuckelaer et al., 2013). Further, test-takers still have to contrast the two opposing statements within one comparison, which means they cannot fully endorse all desirable statements.

Faking in Graded Paired Comparisons

Whether GPCs have the potential to actually be more faking resistant in high-stakes situations has not, to our knowledge, been empirically tested. In the following, we present recent findings on the fakeability of discrete forced-choice formats as well as Likert scales. Based on this, we discuss which faking effects can be expected for GPCs, especially in comparison to Likert scales as a standard method of questionnaire-based personality measurement. Further, we will propose two important boundary conditions of faking in GPCs, namely the faking intention of the individual respondent and the difference in social desirability of both items involved in a given GPC.

Overall Faking Effects

Are people able to deliberately distort their answers in personality questionnaires, for instance, in order to present themselves more favorably when applying for a job? The simple answer to this question is yes. According to meta-analyses, respondents can improve their scores substantially if they are instructed to do so (meta-analytically by $d = .73$ in Edens & Arthur, 2000, and, depending on the trait the effect size ranges between $d = .47$ for agreeableness and $d = .93$ for emotional stability in Viswesvaran & Ones, 1999). In situations where participants were faced with real-life incentives to distort (e.g., money or a job), the distortion seems to be somewhat weaker ($d = .30$ in a meta-analysis by Edens & Arthur, 2000; see also Birkeland et al., 2006; Martínez & Salgado, 2021).

We are not aware of any empirical evidence regarding the specific fakeability of GPCs in personality assessment. However, since the theoretical rationale for GPCs' potential to resist faking is the same as for discrete forced-choice formats, we will review the results of the corresponding research. Concerning the effects of faking on discrete forced-choice measures, recent meta-analyses show that respondents do alter their scores. Meta-analytic effect sizes for the overall mean score inflation between the honest and the faking conditions are $d = .06$ (Cao & Drasgow, 2019) and $d = .41$, respectively (Speer et al., 2023; see also Martínez & Salgado, 2021). Thus, test-takers can inflate their scores in both Likert and forced-choice questionnaires. Because essentially the same mechanism is supposed to inhibit response distortion in forced-choice items and GPCs (i.e., one item's gain is the alternative item's loss), we assume that a certain amount of score inflation will also be found in GPCs:

> **Hypothesis 1:** Both rating scales and graded paired comparisons are fakeable. The average trait scores are higher (in the socially desired direction) when respondents try to distort them than when they answer honestly.

However, the meta-analytic faking effects are smaller for forced-choice formats ($d = .41$) than for Likert scales ($d = .75$; Speer et al., 2023, see also Cao & Drasgow). In addition to meta-analyses, which consider the fakeability of one response format (Likert vs. forced-choice) in isolation, individual studies that have drawn a direct comparison can also be used to assess the relative susceptibility to faking. In fact, the vast majority of these studies found faking-induced score inflation to be smaller for forced-choice scales (Christiansen et al., 2005; Huber, 2017; Jackson et al., 2000; Lee et al., 2019; Pavlov et al., 2019; Vasilopoulos et al., 2006); but see Heggestad et al. (2006) for contrary results.

Thus, although test-takers can elevate their mean trait scores in forced-choice questionnaires when motivated to do so, the magnitude is significantly smaller than that which can be obtained when using Likert scales. As GPCs adopt the crucial forced-choice characteristic of contrasting different statements, they should also make it harder to endorse all desirable items, thus impeding faking. We therefore expect:

> **Hypothesis 2:** Graded paired comparisons exhibit a higher resistance to faking attempts than Likert scales.

When analyzing such faking effects, two different approaches can be taken: The first is mean score inflation, as described in the previous section; this is calculated by subtracting the honest score from the faked score and then dividing the result by the standard deviation of the honest scores. The second approach is to assess the correlation

between honest and faked trait scores (Pavlov et al., 2019). In this study, we will test fakeability with both approaches. As we have already discussed, we expect lower mean score inflation in the GPC condition.

Hypothesis 2a: The mean score inflation between honest and faked answers is lower in graded paired comparisons than in Likert scales.

The mean score inflation approach is most suitable for assessing fakeability in questionnaires in which a fixed threshold must be reached, e.g., in order to advance to the next round of an application procedure. However, in many selection processes, absolute values are irrelevant, and instead, the best candidates relative to all applicants are selected. For such situations, the correlation-based approach shows how well the scores from a faking condition can mirror the honest scores. It is sensitive to relative changes of trait scores between respondents, and—unlike the mean score inflation approach—it does not implicitly assume constant score shifts for all respondents. Therefore, we use the correlation between scores of the honest and faking condition as a second measure in which a potentially enhanced resistance of GPCs to faking attempts (as addressed in Hypothesis 2) should be manifested. In this respect, findings for the discrete forced-choice format are mixed. It is particularly important for potential faking in personnel selection, since faking in this context would be closely linked to changes in applicants' rank order and, therefore, is directly relevant to selection decisions. While in one study forced-choice scores from an applicant condition explained more "true" trait variance than their Likert scale counterparts (Christiansen et al., 2005), other studies have failed to consistently show this supposed superiority of discrete forced-choice scales (Guan, 2015; Heggestad et al., 2006; Pavlov et al., 2019). This might be a consequence of reliability issues associated with discrete forced-choice measures (Guan, 2015; Pavlov et al., 2019) or for the study by Heggestad and colleagues, an artifact of ipsative scoring methods. Importantly, reliability issues could be counteracted by the GPC format, and ipsative scoring artifacts could be avoided by scoring responses with Thurstonian IRT models as this would facilitate normative trait score estimates. Furthermore, in the study by Heggestad and colleagues, honest scores were obtained exclusively with a Likert scale, resulting in a common method bias in favor of Likert scales. Thus, we assume that GPCs can preserve more of the test-takers "true" variance than Likert scales can in a study design that considers the previous shortcomings by collecting data for each of the two test formats under both honest and faking conditions while analyzing the GPCs with Thurstonian IRT models thus yielding normative scores:

Hypothesis 2b: The correlation of honest and faked trait scores will be higher in graded paired comparisons compared with Likert scales.

The Role of Individual Faking Intention

As stated before, respondents differ in how much they fake their answers. In personnel selection contexts, estimates suggest that about 40 to 60% of applicants actually alter their scores (Donovan et al., 2003; Griffin & Wilson, 2012; Griffith et al., 2007). Further, there is considerable variance in the extent to which respondents fake their answers (Rosse et al., 1998). Even in studies where participants are asked to fake as much as possible, test takers differ in the degree to which they fake (e.g., McFarland & Ryan, 2000; Pavlov, 2015). Individuals with a higher will to present themselves positively will change their answers more between honest and application condition. As a consequence, the association of trait scores from the honest and faking condition will be weaker with increasing intention to fake. This moderating effect has recently been demonstrated for discrete forced-choice response formats (Pavlov, 2015; Pavlov et al., 2019). Likewise, we assume that the association of honest and faked trait scores (Hypothesis 2b) will be influenced by the individual intention to fake for both the Likert and the GPC format:

Hypothesis 3: As the individual faking intention increases, the association of honest and faked scores will decrease across the two questionnaire formats.

While the theory on response distortion and the empirical findings for discrete forced-choice tasks are in agreement here, another observation calls for more research: In the Pavlov et al. (2019) study, the association of honest and faked trait scores decreased with an increasing individual faking intention for both the Likert and the forced-choice format to an equal extent. To explain this, the authors mention the lower reliability of forced-choice measurements or the fact that manipulations of one trait indicator (i.e., item) in forced-choice inevitably lead to changes in another trait's indicator (i.e., item), which could not be distinguished in their design (Pavlov et al., 2019). The increased reliability of GPCs (Brown & Maydeu-Olivares, 2018) might allow one to test whether the empirical results are in line with the previous theory on a reduced fakeability of forced-choice formats when the reliabilities of both formats are comparable. In line with the assumption that forced-choice techniques can resist even stronger faking attempts, we hypothesize:

Hypothesis 4: The association of honest and faked trait scores will be less affected by the individual faking intention in the GPC format compared with the Likert format.

Social Desirability of Items

Based on the assumptions that GPCs are also fakeable up to a certain degree and that subjects can intentionally control their faking behavior, the question arises how the process of response distortion exactly takes place. Which specific item properties are used to decide on the shift from an honest answer to distorted responses? We assume that respondents prefer more socially desirable items over less socially desirable ones. This assumption emerges from the mechanism by which forced-choice techniques are postulated to reduce faking: by forcing choice between equally socially desirable items. Now one could argue that social desirability matching should be a prerequisite for forced-choice tests and, therefore, unequal desirabilities should not occur. In fact, however, it is virtually impossible that all items of a forced-choice questionnaire that are compared with each other are exactly equally socially desirable. Moreover, many scholars do not match their items for social desirability at all (Cao & Drasgow, 2019). This tendency is amplified by the requirements of Thurstonian IRT models, whose reliability critically depends on the inclusion of unequally keyed items in the questionnaires (Brown & Maydeu-Olivares, 2011; Bürkner et al., 2019; Schulte et al., 2020). Unequally keyed item blocks contain positively and negatively keyed items that are unlikely equally desirable. We believe that this basic assumption about the functioning of forced-choice, which has so far hardly been empirically tested, should be examined more closely, as it has fundamental implications for the test design and the usefulness of the Thurstonian IRT approach. The impact of desirability differences on faking behavior can be tested more efficiently with GPCs than with discrete forced-choice formats, because responses shift on a more fine-grained ordinal scale, such that they can even detect effects of small desirability differences. Thus, we will test the following hypothesis:

> **Hypothesis 5:** The mean score inflation for a given graded paired comparison is positively correlated with the difference in perceived desirability of its two statements.

The hypotheses, the method, and the data analysis procedure of this study were specified prior to the start of data collection. For the preregistration form, see https://osf.io/8emj7.

Method

Participants

Based on a power analysis, the intended sample size was 618 observations, giving an expected effect size of $d = .16$ for the mean score inflation in forced-choice questionnaires (see meta-analysis by Cao & Drasgow, 2019), an α error of .05, a power of .80, and equal sample sizes in all conditions. Simulations show that sample sizes of $N = 300$ are sufficient to estimate Thurstonian IRT models for GPCs (Lingel et al., 2022).

Participants were recruited via an online panel of professionals (*PsyWeb*), social media channels (Facebook, Xing, LinkedIn, Instagram), the online platform *SurveyCircle*, and the student participant pools at the universities of Münster and Ulm (Germany). The data collection took place between September 2019 and February 2020.

Of the original 591 participants, seven were excluded because they were younger than 18 years, one person's [study language, German] skills were insufficient, and five individuals wanted to exclude their data from the analysis after they had completed the study. Moreover, four people were excluded due to implausible short response times in combination with extremely monotonous response patterns. Furthermore, one person was excluded solely due to a very short overall response time. All included participants provided data on all study variables.

Thus, the final sample size was $N = 573$ (78% female). Participants' ages ranged from 18 to 56 ($M = 27.50$, $SD = 9.93$). In our sample, the majority (67%) were students, 20% were employed full-time in a large variety of jobs, and 9% were part-time employees, while 4% were unemployed. The highest educational level was a secondary school certificate for 3%, a university entrance qualification for 61%, and a university degree for 36% of the participants. As an incentive, participants were offered feedback on their ability to fake in applicant settings and psychology students could earn course credits. Additionally, the four "best" applicants were rewarded with 50 € each.

Design

We conducted an online experiment with a 2×2 mixed design. Participants completed a GPC personality measure ($n = 283$) or the Likert version of the same questionnaire ($n = 290$). We used a between-subject design for this factor to ensure sufficient attention of the participants over the entire duration of the online experiment. After the participants had been asked to answer honestly, we presented a job advertisement for the position of a project manager. The participants were then asked to answer the same questionnaire again, but in a way that maximized their chances of getting the job. We used a within-subject design for the honest/faking factor, as several of our hypotheses had to be tested with regression analysis. Additionally, meta-analyses show that this design can detect faking effects better (Edens & Arthur, 2000; Viswesvaran & Ones, 1999). Respondents were randomly assigned to the two response format

conditions, with equal group sizes being enforced with the last participants. In the end of the applicant condition, respondents were asked how much they tried to fake their answers (referred to as individual faking intention in the following). Finally, both groups rated the desirability of the traits that had been measured and the personality items that had been used.

Materials

We used work-related personality items developed for a personnel selection test in a large German governmental organization. They measured the four dimensions of the Big 5 that predict work performance for the main job profile in this organization, namely emotional stability, extraversion, agreeableness, and conscientiousness.

Likert Personality Questionnaire

The questionnaire in the Likert condition consisted of 42 self-describing statements, 10 each measuring extraversion and agreeableness and 11 each measuring neuroticism and conscientiousness. Participants indicated their level of agreement on a 7-point Likert scale from 1 (*do not agree at all*) to 7 (*fully agree*). Empirical reliabilities are reported in Table 1 and can be interpreted as high. Comparisons of measurement accuracy between conditions should be made on the basis of scale-independent reliability estimates and not on the basis of SEs, as the variance of factor scores differs between conditions (see Table 5 in the online supplement).

GPC Personality Questionnaire

The questionnaire in the GPC condition consisted of the same items as the Likert version. All items from the Likert version and only these were used for the GPC version of the questionnaire. However, in the GPC version, some items were used in several item pairs to yield sufficient and with regard to the Likert version similar levels of reliability. In sum, the questionnaire in the GPC condition consisted of 119 item pairs (95 two-dimensional, 24 unidimensional) of which 48 (two-dimensional) GPCs were unequally keyed. Participants were asked to express the degree to which they prefer one of the two statements within a GPC over the other on a 9-point rating scale ranging from 1 (*the statement on the left describes me very much better*) to 9 (*the statement on the right describes me very much better*). Indicators of measurement precision are reported in Table 1. The GPC reliabilities were on the same level or slightly higher, and the SEs were equal or lower than those for Likert scales. However, this comparison should be interpreted with caution, because GPC scales were longer and included more items than Likert scales in our study. Nevertheless, it can be stated that the GPCs achieved excellent reliability.

The item pairs were taken from an existing test and were initially not matched for social desirability. The main analysis and hypotheses tests are all based on this long questionnaire with both equally and unequally (i.e., mixed) keyed items. However, for additional explorative analysis, we conducted all our hypotheses tests again for two more questionnaires which were formed from subgroups of items from the full questionnaire described above. First, the unequally keyed items were excluded, resulting in a questionnaire of 71

Table 1 Reliabilities for Likert items and graded paired comparisons

	Reliability			*RMSE*		
	Overall	Honest	Faking	Overall	Honest	Faking
Likert						
Emotional stability	.96	.92	.91	.20	.30	.40
Extraversion	.95	.93	.87	.24	.28	.37
Agreeableness	.88	.85	.87	.35	.40	.45
Conscientiousness	.95	.92	.83	.25	.31	.61
Graded Paired Comparisons						
Emotional stability	.97	.94	.93	.17	.24	.27
Extraversion	.97	.96	.93	.19	.21	.23
Agreeableness	.93	.90	.93	.26	.32	.34
Conscientiousness	.97	.95	.94	.17	.23	.25

RMSE = root mean squared individual measurement error. Reliabilities for both formats were calculated with the empirical reliability function from the R package *mirt* (Chalmers, 2012). They represent the proportion of estimated true variance on the sum of the estimates for the true variance and the squared mean standard error. The overall estimates for reliability and *RMSE*s are based on honest and faked scores of all participants, which means that each person is included twice in these calculations. The *RMSE*s in the honest and faking condition are scaled with the same *SD* as the corresponding factor scores and can be interpreted on their scale

equally keyed item pairs. Second, we exclusively analyzed those 50 item pairs with the lowest social desirability difference (for details on social desirability measurement, see below). The desirability difference within an item pair of this questionnaire was less than or equal to 1.17 measured on a 7-point scale, and all of these item pairs were equally keyed. Data for these analyses were taken from the complete questionnaire version. Reliabilities over both the honest and the faking condition were sufficient to excellent for all traits in both additional questionnaire versions ranging from 0.76 to 0.91. The mean standardized factor loading[1] was $M = 0.72$ ($SD = 0.12$) for the complete questionnaire, which is a similar level compared with the CFA-based loadings of Likert items of $M = 0.73$ ($SD = 0.17$). Factor loadings for the equally keyed GPC questionnaire were $M = 0.84$ ($SD = 0.06$) and $M = 0.80$ ($SD = 0.09$) for the GPC questionnaire with item pairs matched for social desirability.

Job Advertisement

A job advertisement for the position of a project manager was used to create a realistic context for the faking condition that is similarly appealing for persons with varying professional backgrounds. It consisted of a general task description and a desired personality/skill profile. The profile listed four general competencies, each corresponding to one of the measured dimensions (e.g., for agreeableness: *You are cooperative and sensitive to the needs of your team members*). The job description was designed in such a way that the participants would perceive all measured characteristics as equally desirable for the job described.

Individual Faking Intention

As proposed by Pavlov (2015), the individual intention to fake was measured by asking participants to what extent they faked their answers in order to present themselves positively with respect to the position of a project manager. The item used an 11-point rating scale ranging from 0 (*I did not distort my answers at all*) to 10 (*I distorted my answers strongly*). In all regression analyses, faking intention was z -standardized across both formats.

Perceived Social Desirability

We measured the social desirability of all 42 single items and the 4 global traits on a 7-point scale ranging from 1 (*very undesirable*) to 7 (*very desirable*). Participants were instructed to rate the desirability with respect to the position of a project manager.

Trait Score Estimation and Standardization

In this section, we briefly describe the estimation procedures for participants' trait scores in the Likert and GPC conditions as well as the standardization procedures, which are essential for interpreting the results.

Trait Score Estimation

We estimated GPC trait scores (person parameters) with Thurstonian IRT models for graded preference data (Brown & Maydeu-Olivares, 2011, 2018). See also Bürkner (2022) for a more detailed model description.

For parameter estimation, we used the *thurstonianIRT* package (Bürkner, 2019) in R (R Core Team, 2020). Within this package, we specified an ordinal model using the Bayesian software Stan (Carpenter et al., 2017) as the underlying engine and the EAP estimator. For all model parameters, we used the default priors from the *thurstonianIRT* package, which are weakly informative and do not change the estimates compared to frequentist software (Bürkner et al., 2019). We estimated factor scores from the honest and faking condition in a common model to estimate both types of scores on the same scale. Only to test whether the assumption of identical item parameters in both conditions affects the results notably, we estimated one model for each of the two conditions (Zhang et al., 2020), resulting in highly correlated factor loadings in both models ($r = .98$). All further analyses are based on the joint model. Two participants reached extreme values on several traits and were excluded from factor score-based regression analyses (Hypotheses 1 to 4) without consequences for acceptance or rejection of hypotheses.

To keep GPC and Likert scores as comparable as possible, trait scores for the Likert condition were estimated using a multidimensional graded response IRT model (Samejima, 1969), which is ordinal as well. As software, we used the R package *mirt* (Chalmers, 2012) with the Metropolis–Hastings Robbins-Monro (MHRM) algorithm and MAP factor scores. Sensitivity analyses showed no differences between the selected MHRM estimator and other methods for multidimensional models, namely Monte Carlo Expectation Maximization and Quasi Monte Carlo Expectation Maximization. The estimates of all methods correlated to $r = .99$ or higher.

Trait Score Standardization

To facilitate the interpretation of the regression models, we standardized all trait scores with the mean and standard

[1] Factor loadings of GPCs were standardized by $\lambda_{standardized} = \frac{\lambda_{unstandardized}}{\sqrt{\lambda^2_{unstandardized}+\psi^2}}$, where λ is the factor loading and ψ is the item uniqueness (i.e., the error). Means are based on absolute values of factor loadings.

deviation of the honest scores (within each questionnaire format). As a consequence, each trait score can be interpreted as the difference from the average honest score in honest *SD*s. For example, if person A has a faking (i.e., applicant) score of 2 on extraversion, this means that this score is two honest *SD*s above the average honest score on extraversion. Likewise, Likert and GPC honest scores will have a mean of zero and a standard deviation of one, while the mean of the faked scores will reflect the average difference of Likert/GPC faked scores from the average Likert/GPC honest score in honest *SD*s. Therefore, the means of the faked scores can be interpreted as Cohen's *d* effect sizes.

The instructions, data, the R code for all analyses, a code book, and an online supplement with additional results are available on OSF (https://osf.io/fx4yz/).

Results

As intended by the experimental manipulation (job profile), all traits were perceived as highly desirable for the position in the job advertisement ($M = 6.61$, $SD = 0.84$ for emotional stability, $M = 6.18$, $SD = 0.95$ for extraversion, $M = 5.73$, $SD = 1.17$ for agreeableness, and $M = 6.82$, $SD = 0.59$ for conscientiousness; all measured on a 7-point scale). Agreeableness had comparatively lower but still high trait desirability ratings. The faking instruction within the applicant scenario combined with the financial incentive also appears to have worked successfully: The individual faking intention was relatively high, with an average value of 7.71 ($SD = 2.30$) in the Likert and 7.73 ($SD = 2.53$) in the GPC condition, both measured on a scale from 1 to 11. The faking intention did not significantly differ between groups ($\Delta M = 0.02$, 95% CI [−0.37, 0.42], $t(569.23) = 0.10$, $p = .920$). The effects of demographic variables on faking strength were negligible (see Table 8 in the online supplement). In this regard, no differences were found between students and professionals either. Unless mentioned otherwise, all analyses refer to the complete version of the GPC questionnaire. The results for the other two versions (equally keyed and matched for social desirability) do not differ substantially. They are addressed at the end of the results section.

Table 2 shows the inter-correlations of trait estimates for GPCs (lower triangle) and Likert scales (upper triangle) for the honest and faking condition. GPC inter-trait correlation estimates are similar but not equal when compared with their Likert counterparts. Inter-trait correlations are higher in the faking condition than in the honest condition, which is a typical effect in application situations and indicates that an ideal applicant factor has formed in both formats.

Due to the inter-correlations of traits (Table 2) and the heterogeneity of variances, we tested all hypotheses with multivariate linear models that account for heterogeneous error variances across the range of predictors. We estimated all multivariate models with the R package *brms* for Bayesian regression modeling (Bürkner, 2017, 2018) based on Stan (Carpenter et al., 2017). See the OSF online supplement for model equations. The regression coefficients of these models can be interpreted like those of single multiple regressions, but credible intervals take into account the inter-correlations of traits. In the Bayesian analyses, we consider as significant those results where the value of the null hypothesis is outside the 95% credible interval. Since the methods mentioned above model the data better than the pre-registered tests for Hypotheses 1 and 2 (*t* tests and bivariate correlations) but at the same time maintain the basic idea of the pre-registered methods, we decided to deviate from the original analysis plan in this respect. We report results for the original analysis plan in an online supplement on OSF. The adjustments to the analysis plan had no effect on the acceptance or rejection of hypotheses. Our analysis approach is based on the regression-based moderation framework suggested by Pavlov et al. (2019).

Table 2 Correlation matrix for GPC (lower triangle) and Likert (upper triangle) factor scores

	1	2	3	4	5	6	7	8
1. Emotional Stability (honest)	-	.27***	.10	.38***	.16**	.02	.02	.27***
2. Extraversion (honest)	.59***	-	.17**	.34***	.06	.23***	.17**	.10
3. Agreeableness (honest)	.20***	.15*	-	.16**	.14*	.19**	.30***	.17**
4. Conscientiousness (honest)	.39***	.25***	.30***	-	.17**	.06	.13*	.38***
5. Emotional Stability (fake)	.04	.10	.28***	.19**	-	.73***	.48***	.34***
6. Extraversion (fake)	-.06	.09	.28***	.15*	.82***	-	.51***	.16**
7. Agreeableness (fake)	.04	.15*	.39***	.22***	.64***	.59***	-	.16**
8. Conscientiousness (fake)	.02	.09	.29***	.18**	.82***	.74***	.65***	-

Honest = honest condition; *fake* = faking condition. Upper triangle shows Likert, and lower triangle shows GPC factor score correlations

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 3 Multivariate regression results with the condition as the predictors for trait scores

	Intercept (honest)	Slope (faking)
Likert		
Emotional stability	0.00 [−0.13, 0.14]	2.67 [2.49, 2.85]
Extraversion	0.00 [−0.13, 0.14]	2.00 [1.83, 2.17]
Agreeableness	0.00 [−0.13, 0.12]	0.97 [0.78, 1.15]
Conscientiousness	0.00 [−0.14, 0.15]	3.11 [2.91, 3.30]
Graded Paired Comparison		
Emotional stability	−0.02 [−0.16, 0.13]	2.15 [1.97, 2.31]
Extraversion	−0.02 [−0.15, 0.12]	1.51 [1.35, 1.67]
Agreeableness	−0.03 [−0.15, 0.09]	1.16 [0.99, 1.34]
Conscientiousness	−0.03 [−0.16, 0.11]	2.04 [1.88, 2.21]

Faking = factor score in faking condition. Condition is dummy coded with honest = 0 and faking = 1. Each column reports regression coefficients followed by the corresponding 95% credible interval [in brackets]. Results are based on separate models for each questionnaire format (Likert and GPC). For results based on classical test theory scoring of Likert scales, see the online supplement

We now report the results of our main analyses, which address the question of whether and under what conditions GPCs are faked, and how they compare to Likert scales in this regard. Hypothesis 1 stated that average trait scores are higher when respondents try to distort them than when they answer honestly. Table 3 shows the model results for the regression of trait scores on the condition (dummy coded with honest = 0 and faking = 1). The intercepts represent the predicted values for the honest condition, which are zero for all traits due to the standardization procedure. The slope coefficients represent the predicted values for the faking condition and can be interpreted as Cohen's *d* effect sizes, that is, they represent the mean difference between the honest and the faking condition with the standard deviation of the honest condition as units. Across all traits, effects for both formats are large and 95% CIs for slope parameters (i.e., the differences between honest and faking condition) do not include zero. Thus, results support Hypothesis 1, that is, participants were able to fake both questionnaire types when they were instructed to do so.

Hypothesis 2a stated that GPC scores are less inflated than Likert scores. To test this hypothesis, we predicted the faking scores, which can be interpreted as the scores' inflation from the honest to the faking condition, by the questionnaire format (dummy coded with Likert = 0 and GPC = 1) while controlling for honest scores and the interaction of honest scores and questionnaire format. At an honest score of 0, the score inflation was significantly lower for GPCs than for Likert scales in three out of four traits (see credible intervals for format coefficients in Table 4). Consequently, respondents seem less able to raise their scores on GPCs than on Likert scales (as predicted in Hypothesis 2a).

Hypothesis 2b stated that the correlation of honest and faked scores would be higher in GPCs compared with Likert scales. We tested this on the basis of the interaction term for the honest score and the questionnaire format (see Table 4). Although all interaction terms are significantly different from 0 for all traits but agreeableness, their signs indicate an effect contrary to the hypothesis and what we would expect based on the results for the mean score inflation: The association of honest and faked scores is higher for Likert scales than for GPCs. This is consistent with the results for trait-wise direct comparisons of the honest–faking correlations for Likert scales and honest–faking correlations for GPCs (see the OSF online supplement). They also do not support Hypothesis 2b but suggest that in GPCs, honest and fake ratings are less correlated than in Likert scales.

The Role of Individual Faking Intention

Next, we analyzed the impact of participants' individual faking behavior as a potential boundary condition of the observed faking effects. We hypothesized that the association of participants' honest and faked scores will decrease the more they distort their responses in the faking condition, regardless of the test format (Hypothesis 3). To investigate this hypothesis, we ran standardized multivariate multiple regression models for Likert and GPC formats. Faked trait scores were regressed on corresponding honest scores, individual faking intention, and the interaction of these variables. Assuming a positive regression coefficient for honest scores, the interaction term should have

Table 4 Multivariate multiple regression results with the interaction of honest scores and format as predictors for faked scores

	Intercept	Honest	Format	Honest*Format
Emotional Stability	2.67 [2.54, 2.80]	0.21 [0.14, 0.29]	−0.56 [−0.73, −0.39]	−0.12 [−0.22, −0.02]
Extraversion	2.01 [1.90, 2.11]	0.19 [0.13, 0.25]	−0.51 [−0.65, −0.38]	−0.13 [−0.22, −0.04]
Agreeableness	0.97 [0.83, 1.10]	0.28 [0.16, 0.39]	0.16 [−0.03, 0.35]	0.01 [−0.16, 0.18]
Conscientiousness	3.11 [2.96, 3.25]	0.17 [0.09, 0.25]	−1.09 [−1.28, −0.91]	−0.17 [−0.29, −0.07]

Honest = honest scores; each column reports regression coefficients followed by the corresponding 95% credible interval [in brackets]. Format is dummy coded with Likert = 0 and graded paired comparisons = 1. Honest*Format is the interaction effect of these two variables

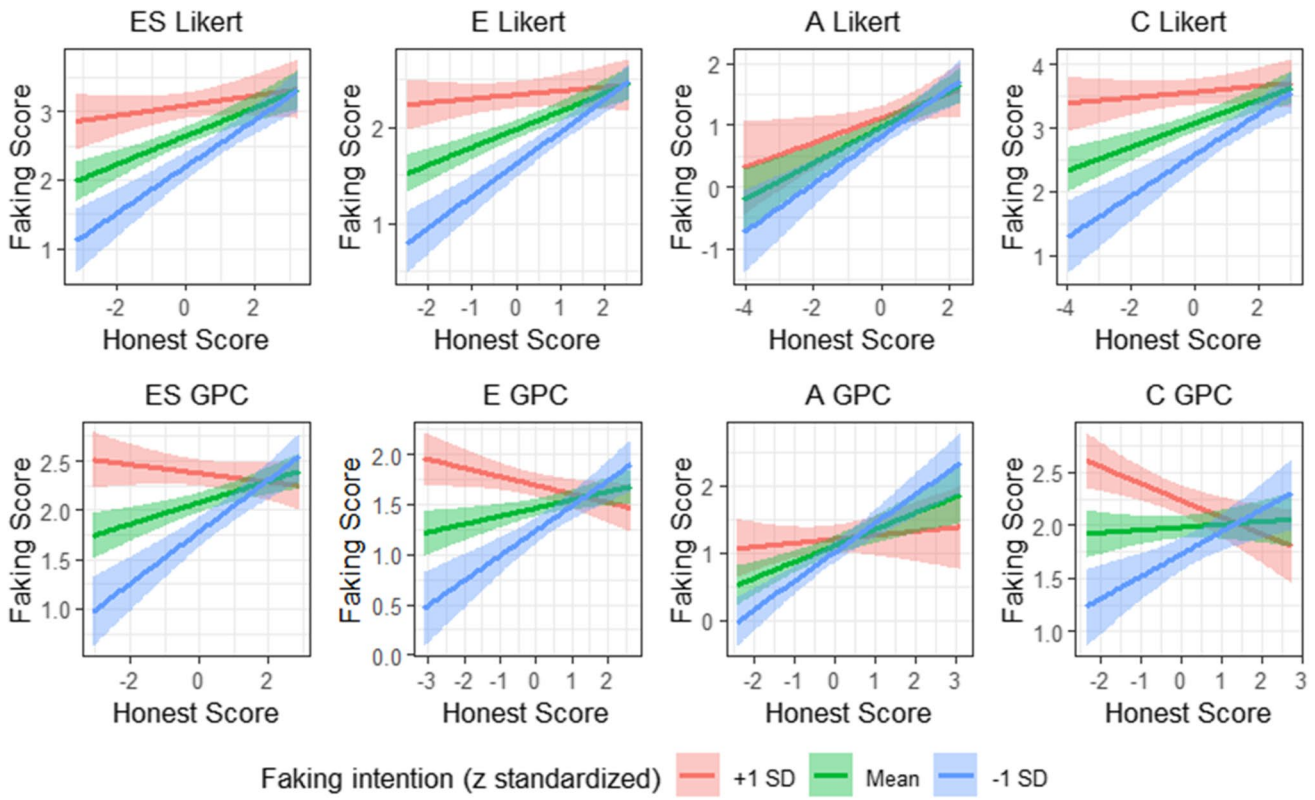
a negative sign, thus reducing the regression weight of honest scores as faking intention increases.

The interaction coefficients were negative in all cases and significant for all traits but agreeableness in the Likert format and for all traits in the GPC format (see *CI*s for the interaction terms in Table 5). Figure 2 plots the interactions of honest scores and individual faking intention. For Likert scales (upper row of plots), we see that the slope becomes flatter the higher the faking intention becomes, that is, the more respondents try to fake, the less the "true" values are mirrored in their faked responses. The same is true for GPCs (lower row), but here, we see that when the faking intention

Table 5 Multivariate multiple regression results with the interaction of honest scores and individual faking intention as predictors for faked scores

	Intercept	Honest	Faking intention	Honest*faking int.
Likert				
Emotional Stability	2.64 [2.51, 2.77]	0.21 [0.13, 0.29]	0.45 [0.31, 0.59]	−0.13 [−0.21, −0.05]
Extraversion	1.98 [1.88, 2.08]	0.19 [0.13, 0.25]	0.36 [0.26, 0.46]	−0.15 [−0.20, −0.09]
Agreeableness	0.97 [0.83, 1.10]	0.29 [0.17, 0.41]	0.15 [0.01, 0.28]	−0.09 [−0.20, 0.02]
Conscientiousness	3.07 [2.93, 3.22]	0.18 [0.11, 0.26]	0.50 [0.34, 0.65]	−0.14 [−0.21, −0.07]
Graded Paired Comparison				
Emotional Stability	2.08 [1.97, 2.18]	0.11 [0.05, 0.17]	0.30 [0.20, 0.41]	−0.15 [−0.21, −0.10]
Extraversion	1.46 [1.37, 1.55]	0.08 [0.02, 0.14]	0.23 [0.14, 0.31]	−0.17 [−0.23, −0.10]
Agreeableness	1.10 [0.96, 1.25]	0.24 [0.12, 0.36]	0.10 [−0.02, 0.22]	−0.19 [−0.29, −0.08]
Conscientiousness	1.98 [1.87, 2.09]	0.03 [−0.06, 0.11]	0.26 [0.15, 0.36]	−0.18 [−0.26, −0.11]

Honest = honest scores, *Honest*faking int.* = interaction of honest score and faking intention. Each column reports regression coefficients followed by the corresponding 95% credible interval [in brackets]. Faking intention is *z* transformed



Note. ES = emotional stability; E = extraversion; A = agreeableness; C = conscientiousness; GPC = graded paired comparison condition. Shaded areas represent 95% credible intervals

Fig. 2 Interaction plots for Hypothesis 2

is high (i.e., one *SD* above the mean), the association of honest and faked scores is actually mostly negative (see negative slopes in three out of four traits). We will return to this point in the explorative results section.

The differences between the Likert and the GPC format concerning the effects of faking intention were addressed by Hypothesis 4: We hypothesized that the individual faking intention would affect the association between honest and faked scores in GPCs less than in the Likert format. To examine this issue, we again ran multivariate multiple regression analyses regressing faked trait scores on the format (Likert vs. GPC), the honest scores, individual faking intention, and all possible interactions of these variables.

The relevant statistics for Hypothesis 4 are the regression coefficients of the threefold interaction *Format* × *Honest Scores* × *Faking Intention*. As shown in the section on Hypothesis 3, the association of honest and faked scores—that is, the regression weight for the predictor *Honest Scores*—becomes smaller as faking intention increases (= negative coefficient for *Honest Scores* × *Faking Intention* interaction term). If this effect is more pronounced in case of the Likert format, the coefficient of the threefold interaction should be of positive magnitude, thus shifting the coefficient of the interaction *Honest Scores* × *Faking Intention* closer to zero when switching from the Likert format (reference group) to the GPC format. However, the threefold interaction effects were very small and non-significant for all four traits (regression coefficients and 95% credible intervals were −0.03 [−0.13, 0.06] for emotional stability, −0.03 [−0.12, 0.06] for extraversion, −0.10 [−0.25, 0.05] for agreeableness, and −0.03 [−0.14, 0.08] for conscientiousness). Thus, Hypothesis 4 is not supported. For all coefficients of the regression model, see the online supplement.

The Role of Perceived Item Desirability

Hypothesis 5 aimed at explaining which item characteristics respondents draw on in GPCs to produce the faking effects observed above. We hypothesized that the mean score inflation for a given GPC is positively associated with the difference in perceived desirability of its two statements. To test this, we computed the difference of perceived desirability for each GPC by subtracting the average perceived desirability of the left statement from the average perceived desirability of the right statement within each GPC. We also calculated the mean score inflation for each GPC by subtracting the raw mean score of a given GPC in the honest condition from its corresponding raw mean score in the faking condition. Therefore, the higher the absolute value of the score inflation, the more the mean for this item has changed between the honest and faking condition.

The correlation between desirability difference of both items in a GPC and score inflation in this GPC is $r = .94$, 95% confidence interval [.91, .96], $t(117) = 29.21$, $p < .001$. Moreover, the intercept of the unstandardized regression of the score inflation on the desirability difference is $b = 0.10$, 95% confidence interval [0.00, 0.21]. Thus, when both items are equally desirable (i.e., their desirability difference is zero), the predicted score inflation is only .10 *SD*s. Results are graphically displayed in Fig. 3. In this figure, red points represent equally keyed GPCs, and turquoise points represent unequally keyed GPCs. Equally keyed GPCs clearly show less absolute score inflation than unequally keyed GPCs, forming nearly perfectly separable classes with very little overlap.

Explorative Analyses

In the following, we report analyses that we performed in knowledge of the results presented above and that were not pre-registered. In this context, we are particularly interested in the counterintuitive results on the interaction of honest scores and individual faking intention in the prediction of faking scores as well as the influence of item desirability on response behavior.

In Fig. 2, GPC honest and faked scores are mostly negatively associated when faking intention is high. We wanted to analyze these effects more closely. Further, we aimed to exclude the possibility that this is a statistical artifact; potential non-linear effects are represented exclusively by the interaction term, meaning that the data was potentially modeled inadequately. Therefore, we allowed for linear and quadratic effects in both predictors and their error variances. Figure 4 describes the same association with the faking intention as a continuous variable and from a different perspective. The color describes the value of the trait score in the faking condition, and the lines run along constant values. Two different types of respondents reached high scores in the faking condition: first, those with a high faking intention and low scores in the honest condition, and second, those with high honest scores but a low faking intention. Those who showed medium levels of both predictors or even high levels of both predictors seem to yield lower scores in the faking condition than those with very high scores in the honest condition and a very low faking intention or vice versa. Even though the patterns are very consistent across the traits, the results have to be interpreted with caution because the trait-wise credibility intervals of the honest scores' quadratic effects all include zero.

We also conducted additional analyses regarding social desirability. Results for Hypothesis 5 suggest that removing the unequally desirable items from the GPC questionnaire might reduce its fakeability. We ran the regression models for Hypotheses 1 to 4 again with the two other versions of the GPC questionnaire, namely the equally keyed version and that with 50 item pairs with the lowest desirability difference. The

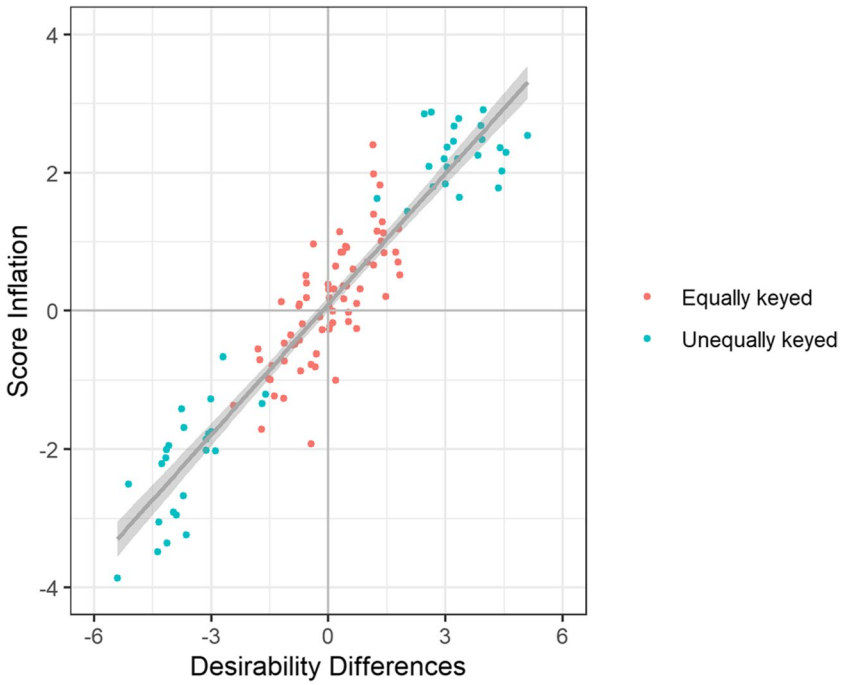
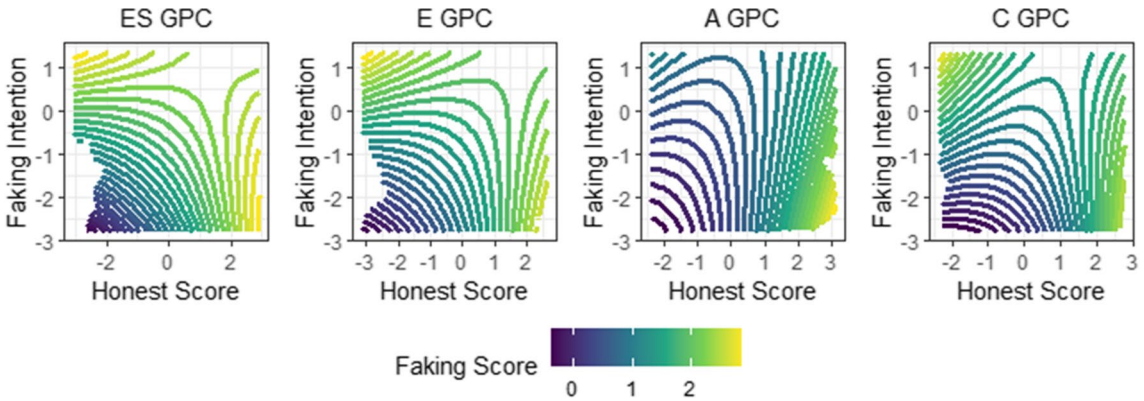


Fig. 3 Scatter plot and unstandardized regression line for the influence of desirability difference on score inflation in GPCs

Likert questionnaire remained unchanged. The results of these analyses are presented in the online supplement. They do not indicate a better faking resistance of the GPC factor scores neither when unequally keyed item pairs are removed nor when items were matched for social desirability.

Discussion

The main purpose of this study was to judge the utility of GPCs for personality assessment in high-stakes situations such as personnel selection. As expected, participants managed to fake both Likert and GPC scales, i.e., the mean trait scores were higher when respondents tried to distort them than when they answered honestly. However, the mean score inflation was lower in GPCs compared with Likert scales. Remarkably, the reduced score inflation did not translate into a closer association of honest and faked trait scores. Even though the GPC scores are less inflated, they do not seem to capture the real differences between respondents more adequately in situations where they are motivated to distort. Thus, the assertion that GPCs would exhibit a higher faking resistance was only partially supported. One factor on the respondent side that influences the association of honest and faked scores is the individual faking intention. In both questionnaire types, scores can



Note. ES = emotional stability; E = extraversion; A = agreeableness; C = conscientiousness; GPC = graded paired comparison condition. Shaded areas represent 95% credible intervals

Fig. 4 GPC interaction plots for Hypothesis 2 with continuous variables and quadratic effects

be manipulated if the respondent tries to do so. In Likert scales, the scores in the faking condition increased with higher faking intentions and higher scores in the honest condition. If one of the two (faking intention or honest trait score) was highly pronounced, the other showed little additional effects. In contrast, in GPCs, trait scores in the faking condition were the highest when either the honest trait scores or the faking intention was high and the other variable was low. The association of honest and faked trait scores was even mostly negative when the faking intention was high. Moreover, the association of these variables in GPC questionnaires could be non-linear, but our evidence is not conclusive in this respect. Finally, an extremely important questionnaire feature that determines the response inflation in a given GPC is the desirability difference between the two items involved. This desirability difference is much larger in unequally keyed item pairs than in equally keyed ones.

Theoretical Implications

The results confirm and expand our knowledge about the forced-choice format in general and GPCs in particular. The meta-analytic finding that forced-choice scores are inflated when respondents are motivated to fake but less so than in Likert questionnaires (Cao & Drasgow, 2019; Speer et al., 2023) seems to apply to GPCs as well. The found effect sizes were high, and—if no further precautions are taken to avoid faking—the association of honest and faked scores is low. Compared with most other faking studies and meta-analytic estimates, our effect sizes are relatively large for both GPC and Likert questionnaires. Note that in faking studies, the effect sizes depend on several design decisions. Faking effects are typically higher in instructionally induced designs as ours than in studies with real-life motivational distortion (Cao & Drasgow, 2019; Edens & Arthur, 2000). This is because designs with faking instructions directly manipulate the motivation to distort which is more heterogeneous in real application contexts (i.e., some applicants do not distort at all; Donovan et al., 2003; Griffin & Wilson, 2012; Griffith et al., 2007). By omitting the null effects of respondents without faking motivation, we isolated the effects of respondents trying to improve their scores. Thus, our results are representative for people who are willing to show high levels of impression management in order to be selected. These are exactly those applicants that faking reduction methods such as forced-choice were developed for. For such conditions, our results suggest that GPCs compared with Likert scales can reduce score inflation to some extent, but still, they raise questions about the construct validity of personality tests when administered to faking-motivated respondents. Specifically, meta-analyses (e.g., Barrick & Mount, 1991; Schmidt & Hunter, 1998) may overestimate the actual predictive validity of personality tests for this part of the applicant population: A closer look at corresponding studies shows that they rely on participants already employed by a company, meaning they may have a lower faking motivation than job applicants.

The lower correlation between honest and faked trait scores for forced-choice formats compared with Likert items has previously been observed for discrete forced-choice formats (Guan, 2015; Heggestad et al., 2006; Pavlov et al., 2019). However, considering our study's design and analysis strategy, we can rule out alternative interpretations mentioned in earlier studies. Particularly, in our study, neither the ipsative classical test theory scoring method (Heggestad et al., 2006) nor the lower reliability of discrete forced-choice formats (Guan, 2015; Pavlov et al., 2019) are valid explanations. Instead, respondents with medium or high honest scores reached systematically lower scores in the applicant condition than those who had low true, i.e., honest, scores. The negative association of honest and faked scores in a subpopulation diminishes the overall correlation for the full population. Furthermore, non-linear effects might negatively influence the level of correlation. Together with the diminished correlations between honest and faked scores in the forced-choice questionnaire reported previously (Guan, 2015; Heggestad et al., 2006; Pavlov et al., 2019), these observations point to a severe psychometric weakness of the forced-choice format. We assume that this effect is caused by artifacts of the interdependent nature of the response process and/or format. Each response change in favor of one item inevitably causes a change in the response to another item and, therefore, the corresponding trait score estimate. The consequences of these manipulations are difficult to foresee for respondents and might unintentionally negatively affect other traits' scores that they did not intend to affect.

Our study also aimed to reveal questionnaire properties that respondents use to decide on the shift from an honest answer to distorted responses. In this respect, it seems as if the social desirability is crucial. Desirability differences between items and the score inflation in this item pair are very closely associated. Thus, if the difference between the desirability of two items in a given situation is known, it can be used to predict almost perfectly how much these items are distorted. Moreover, GPCs have the potential to reduce mean response shifts from the honest to the faking condition to an absolute minimum if item desirabilities are exactly matched. In addition, results on raw response level suggest very clearly that unequally keyed item pairs are faked particularly strongly. The responses to these item pairs changed considerably more between honest and faking conditions than responses to equally keyed items did. This is most likely due to the large differences in desirability of these item pairs. The present study provides

empirical support for the assumption that in unequally keyed item pairs, one item typically represents the desired end and one the undesired end of a trait continuum, resulting in large desirability differences (Bürkner et al., 2019; Schulte et al., 2020).

Based on the high score inflation in item pairs with higher social desirability differences (what particularly concerns unequally keyed items), it would have been reasonable to assume that the score inflation in Thustonian IRT factor scores could be reduced by removing these item pairs. However, this was not the case in our explorative analyses. This is counterintuitive on the first view but is in line with the meta-analysis by Cao and Drasgow (2019) which found lower faking effects for normative than for ipsative scores in discrete forced-choice questionnaires. Ipsative scores rely on sums of raw responses (comparable with classical test theory scoring for Likert scales). Thus, changes in raw responses directly translate into changes in trait scores. In contrast, normative scoring methods like Thurstonian IRT (as used in our study) weight each response differently. Items that are strongly faked by almost all respondents contribute little to precise trait score estimation under faking conditions due to the homogeneous responses to these items. We assume that this is the reason why results remained nearly unchanged when the most strongly faked item pairs were removed in explorative analysis.

Another noteworthy observation from the present study relates to the faking differences between traits, especially with regard to agreeableness. In line with meta-analytic results (Speer et al., 2023), traits that were perceived as more desirable seemed to have been faked more strongly. Note that social desirability is context-specific. Items and traits that are highly desirable for one job can be undesirable for another job and vice versa. Accordingly, faking behavior will depend on the job context as well.

Practical Implications

GPCs represent an important extension of forced-choice-based response formats. In particular, they seem to allow for highly reliable trait estimates even with only equally keyed item pairs, which successfully eliminates a major weakness of discrete forced-choice formats (Bürkner et al., 2019; Schulte et al., 2020). In this respect, they are preferable to other discrete forced-choice formats. Regarding their fakeability, it seems that GPCs have similar strengths and weaknesses as have been reported for discrete forced-choice formats. They can reduce the score inflation in job applicants or other respondents who are motivated to fake in high-stakes situations, but for practical applications, they can currently only be recommended to a limited extent. GPCs seem to be less capable than Likert scales of recovering applicants' true aptitude relative to other applicants. To put it simple, GPCs are less helpful than Likert scales in answering the question of whether candidate A is better suited than candidate B in a situation where both are motivated to fake. Thus, these scales should not be used to make important decisions until these problems have been resolved. Further research must show whether this is possible.

When test developers attempt to construct forced-choice questionnaires, they should pay particular attention to the social desirability of the items, as the responses can only be expected to be relatively free of faking effects if the items to be compared are equally socially desirable. Accordingly, questionnaire development should start with an item pool that is large enough to exclude items that cannot be paired based on their desirability. It is reasonable to assume that unequally keyed item pairs/blocks contribute little or nothing to construct-valid trait estimation in high-stakes situations, as they are strongly faked.

On the other hand, since the exclusion of differently socially desirable pairs had no significant influence on the faking effect strength on factor score level, it can also be argued that these item pairs do not cause any harm. This seems to be possible at least for IRT-based analysis methods, since homogeneously answered items can be weighted less due to the low level of information they provide. If one follows this argumentation, however, an exclusion of unequally desired items would still be advisable, because in this case, they would be unnecessary for the factor score estimation and thus extend questionnaire length without any need. Further, our study used an induced faking design which leads to higher and more similar intentions to distort. In practice, not all respondents fake to an equal extent. Some applicants just respond honestly. This leads to more heterogeneous responses which in turn increases the amount of (in part distorted) information in the faked items and thus increases the influence of these item pairs on trait scores. Consequently, the need to exclude these items might be even more pronounced under real applicant conditions.

Limitations and Future Research

Since sufficient levels of reliability can be achieved with GPCs, the most severe remaining weakness of GPCs (and forced-choice formats in general) seems to be their limited ability to recover the true rank order of respondents' trait scores. Further research is necessary to determine why this is the case, and doing so would require gaining a better understanding of the faking processes in forced-choice questionnaires. Although social desirability seems to play a major role here, the exact response processes and their consequences for the estimated factor scores remain largely unknown. Further, the negative effects of high

faking intentions need to be examined in depth, as it would be interesting to learn whether this phenomenon is caused by the cognitive response process, the response format, response strategies specific to this format, or the analysis method.

One question that comes with the construction of GPCs is the use of a middle response category. On the one hand, equal preferences do occur and the middle category gives the chance to indicate this correctly. On the other hand, an even number of response categories could force more (but possibly artificial) differentiation. It is also possible that the middle category is particularly attractive under faking conditions. The advantages and disadvantages of a middle category should, therefore, be examined in terms of how well it is able to capture true trait values.

Finally, as some of our results are explorative, they should be interpreted with caution until successfully replicated.

Conclusion

To our knowledge, this study is the first to investigate the performance of the GPC format under high-stakes (i.e., faking) conditions. The graded response format can provide reliable personality measures and should, thus, be considered as an alternative to discrete forced-choice formats. From a research perspective, it allows for more detailed insights into the general response distortion of forced-choice response formats. Our study also indicates that some important aspects should be considered when designing faking-resistant forced-choice questionnaires. The social desirability of the items—which in turn is extremely closely linked to item keying—is of outstanding importance. Social desirability has a very high impact on response behavior and score inflation. Furthermore, the faking effects we found resemble those of discrete forced-choice formats: forced-choice questionnaires constructed according to current knowledge can reduce score inflation but do not seem to outperform Likert scales in the recovery of applicants' true trait standing. Although the mean score inflation paradigm has previously dominated faking research, especially on forced-choice questionnaires, it does not seem to account for important disadvantages of forced-choice-based formats. If the effects found here are also found for discrete forced-choice formats, which earlier studies suggest (Guan, 2015; Heggestad et al., 2006; Pavlov et al., 2019), then the relative fakeability of forced-choice vs. Likert scales should be revised. To judge whether the issues we found can be resolved by improved questionnaire design and/or alternative analysis methods, a better understanding of the faking process and its effects on trait estimates is necessary.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10869-024-09931-0.

Author Contribution NS and HH contributed to the study idea and design. NS and LK developed the hypotheses, created the questionnaire, determined the evaluation strategy, and collected and analyzed the data. PCB provided helpful support for the data analysis. NS prepared the first draft of the manuscript, taking LKs' master's thesis into account. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. The manuscript is based on the master's thesis of LK and is part of the dissertation of NS.

Funding Open Access funding enabled and organized by Projekt DEAL. This research was partially supported by the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation).

Open Science Information This study was registered on OSF (https://osf.io/8emj7). All data and analysis scripts are available on https://osf.io/fx4yz/. To conduct the presented analyses and create this article, we used the programming language R (Version 4.0.2; R Core Team, 2020) and the R-packages *brms* (Version 2.13.5; Bürkner, 2017, 2018), *corx* (Version 1.0.6.1; Conigrave, 2020), *dplyr* (Version 1.0.1; Wickham et al., 2020), *ggplot2* (Version 3.3.2; Wickham, 2016), *gridExtra* (Version 2.3; Auguie, 2017), *kableExtra* (Version 1.1.0; Zhu, 2019), *lavaan* (Version 0.6.7; Rosseel, 2012), *mirt* (Version 1.32.1; Chalmers, 2012), *MVN* (Version 5.8; Korkmaz et al., 2014), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *rstan* (Version 2.21.2; Stan Development Team, 2020), *thurstonianIRT* (Version 0.11.1; Bürkner, 2019), *tibble* (Version 3.0.3; Müller & Wickham, 2020), *tidyr* (Version 1.1.1; Wickham & Girlich, 2020), *tidyverse* (Version 1.3.0; Wickham et al., 2019).

Declarations

Conflict of Interest The authors declare no competing interests.

References

Agresti, A. (1992). Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 41*(2), 287–297. https://doi.org/10.2307/2347562

Alfaro-Rodriguez, H., O'Mahony, M., & Angulo, O. (2005). Paired preference tests: D' values from Mexican consumers with various response options. *Journal of Sensory Studies, 20*(3), 275–281. https://doi.org/10.1111/j.1745-459X.2005.00018.x

Auguie, B. (2017). gridExtra: Miscellaneous functions for "Grid" Graphics". *R package version*, *2*, 3. https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown*. R package version 0.1.0.9997. https://github.com/crsh/papaja

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

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

Borman, T. C., Dunlop, P. D., Gagné, M., & Neale, M. (2023). Improving reactions to forced-choice personality measures in simulated job application contexts through the satisfaction of psychological needs. *Journal of Business and Psychology*. https://doi.org/10.1007/s10869-023-09876-w

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

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

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Bürkner. (2019). thurstonianIRT: Thurstonian IRT models in R. *Journal of Open Source Software, 4*(42), 1662. https://doi.org/10.21105/joss.01662

Bürkner, P.-C. (2022). On the information obtainable from comparative judgments. *Psychometrika, 87*(4), 1439–1472. https://doi.org/10.1007/s11336-022-09843-z

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

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

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Ridell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51*(5), 292–303. https://doi.org/10.1037/h0057299

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

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

Clemans, W. V. (1966). *An analytical and empirical examination of some properties of ipsative measures [Psychometrika Monograph No. 14]*. Psychometric Society. Retrieved January 15, 2024, from https://www.psychometricsociety.org/sites/main/files/file-attachments/mn14.pdf

Conigrave, J. (2020). *corx: Create and format correlation matrices*. R package version 1.0.6.1. https://CRAN.R-project.org/package=corx

Dalal, D. K., Zhu, X., Rangel, B., Boyce, A. S., & Lobene, E. (2019). Improving applicant reactions to forced-choice personality measurement: Interventions to reduce threats to test takers' self-concepts. *Journal of Business and Psychology, 36*(1), 55–70. https://doi.org/10.1007/s10869-019-09655-6

De Beuckelaer, A., Kampen, J. K., & Van Trijp, J. C. M. (2013). An empirical assessment of the cross-national measurement validity of graded paired comparisons. *Quality & Quantity: International Journal of Methodology, 47*(2), 1063–1076. https://doi.org/10.1007/s11135-011-9583-1

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

Edens, P. S., & Arthur, W. (2000). *A meta-analysis investigating the susceptibility of self-report inventories to distortion* [Poster presentation]. 15th annual conference of the society for industrial and organizational psychology, New Orleans, LA.

Griffin, B., & Wilson, I. G. (2012). Faking good: Self-enhancement in medical school applicants. *Medical Education, 46*(5), 485–490. https://doi.org/10.1111/j.1365-2923.2011.04208.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), 341–355. https://doi.org/10.1108/00483480710731310

Guan, L. (2015). *Personality, faking, and the ability of identify criteria: Can forced choice formats untangle their relationships? [Unpublished master thesis]*. University of Virginia.

Guenole, N., Brown, A., & Cooper, A. J. (2018). Forced-choice assessment of work-related maladaptive personality traits: Preliminary evidence from an application of Thurstonian item response modeling. *Assessment, 25*(4), 513–526. https://doi.org/10.1177/1073191116641181

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. https://doi.org/10.1037/0021-9010.91.1.9

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*(3), 167–184. https://doi.org/10.1037/h0029780

Hontangas, P. M., de la Torre, J., Ponsoda, V., Leenen, I., Morillo, D., & Abad, F. J. (2015). Comparing traditional and IRT scoring of forced-choice tests. *Applied Psychological Measurement, 39*(8), 598–612. https://doi.org/10.1177/0146621615585851

Huber, C. R. (2017). *Faking and the validity of personality tests: Using new faking-resistant measures to study some old questions* [Doctoral dissertation, University of Minnesota]. Retrieved May 17, 2018, from https://conservancy.umn.edu/bitstream/handle/11299/185605/Huber_umn_0130E_17909.pdf?sequence=1&isAllowed=y

Huber, J., & Holbrook, M. B. (1982). Estimating temporal trends in preferences measured by graded paired comparisons. *Journal*

of Business Research, 10(4), 459–473. https://doi.org/10.1016/0148-2963(82)90005-4

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. https://doi.org/10.1207/S15327043HUP1304_3

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*(2), 153–162. https://doi.org/10.1111/j.2044-8325.1988.tb00279.x

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal., 6*(2), 151–162.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing & Health, 25*(4), 295–306. https://doi.org/10.1002/nur.10041

Lee, P., Joo, S.-H., & Lee, S. (2019). Examining stability of personality profile solutions between Likert-type and multidimensional forced choice measure. *Personality and Individual Differences, 142*, 13–20. https://doi.org/10.1016/j.paid.2019.01.022

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 5–55. Retrieved January 15, 2024, from https://legacy.voteview.com/pdf/Likert_1932.pdf

Lingel, H., Bürkner, P.-C., Melchers, K., & Schulte, N. (2022). *Measuring personality when stakes are high: Are graded paired comparisons a more reliable alternative to traditional forced-choice methods?*. PsyArXiv. https://doi.org/10.31234/osf.io/8rt3j

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

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

Müller, K., & Wickham, H. (2020) tibble: Simple data frames. R package version 3.0.3. https://CRAN.R-project.org/package=tibble

Ofir, C. (2004). Reexamining latitude of price acceptability and price thresholds: Predicting basic consumer reaction to price. *Journal of Consumer Research, 30*(4), 612–621. https://doi.org/10.1086/380293

Paulhus, D. L., & Jones, D. N. (2014). Measurement of dark personalities. In G. J. Boyle, D. H. Saklofske, & G. Matthews (Eds.), *Measures of personality and social psychological constructs* (pp. 562–594). Academic Press. https://doi.org/10.1016/B978-0-12-386915-9.00020-6

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). Guilford.

Pavlov, G. (2015). *Intentional response distortion effects on personality scores in simulated personnel assessment settings: A moderation study* [Doctoral dissertation, ie University, Madrid, Spain]. Retrieved March 14, 2018, from https://www.researchgate.net/publication/307599102_Intentional_Response_Distortion_Effects_on_Personality_Scores_in_Simulated_Personnel_Assessment_Settings_A_Moderation_Study

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. https://doi.org/10.1177/1094428117753683

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rasheed, S., & Robie, C. (2023). Faking resistance of a quasi-ipsative RIASEC occupational interest measure. *International Journal of Selection and Assessment, 31*, 321–335. https://doi.org/10.1111/ijsa.12427

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*(4), 634–644. https://doi.org/10.1037/0021-9010.83.4.634

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4, Pt. 2), 100. https://doi.org/10.1007/BF03372160

Saville, P., & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. *Journal of Occupational Psychology, 64*(3), 219–238. https://doi.org/10.1111/j.2044-8325.1991.tb00556.x

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. https://doi.org/10.1037/0033-2909.124.2.262

Scholz, S. W., Meissner, M., & Decker, R. (2010). Measuring consumer preferences for complex products: A compositional approach based on paired comparisons. *Journal of Marketing Research, 47*(4), 685–698. https://doi.org/10.1509/jmkr.47.4.685

Schulte, N., Holling, H., & Bürkner, P.-C. (2020). Can high-dimensional questionnaires resolve the ipsativity issue of forced-choice response formats? *Educational and Psychological Measurement, 81*(2), 262–289. https://doi.org/10.1177/0013164420934861

Skedgel, C. D., Wailoo, A. J., & Akehurst, R. L. (2015). Choosing vs allocating: Discrete choice experiments and constant-sum paired comparisons for the elicitation of societal preferences. *Health Expectations: An International Journal of Public Participation in Health Care & Health Policy, 18*(5), 1227–1240. https://doi.org/10.1111/hex.12098

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

Stan Development Team. (2020). RStan: The R interface to Stan. *R package version*, *2*(21), 2. https://mc-stan.org/

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. https://doi.org/10.1207/s15327043hup1903_1

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

Wang, W.-C., Qiu, X.-L., Chen, C.-W., Ro, S., & Jin, K.-Y. (2017). Item response theory models for ipsative tests with multidimensional pairwise comparison items. *Applied Psychological Measurement, 41*(8), 600–613. https://doi.org/10.1177/0146621617703183

Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349–363). Oxford University Press.

Wetzel, E., & Greiff, S. (2018). The world beyond rating scales: Why we should think more carefully about the response format in

questionnaires. *European Journal of Psychological Assessment, 34*(1), 1–5. https://doi.org/10.1027/1015-5759/a000469

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Wickham, H., & Girlich, M. (2020). tidyr: Tidy messy data. *R package version*, *1*(1), 1. https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., & Müller, K. (2020). dplyr: A grammar of data manipulation. R package version 1.0.1. https://CRAN.R-project.org/package=dplyr

Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Young, A. L. (2018). *Faking resistance of a forced-choice measure of the dark triad* [Doctoral dissertation, North Carolina State University]. Retrieved January 15, 2024, from http://www.lib.ncsu.edu/resolver/1840.20/35649

Zhang, B., Cao, M., Tay, L., Luo, J., & Drasgow, F. (2020). Examining the item response process to personality measures in high-stakes situations: Issues of measurement validity and predictive validity. *Personnel Psychology, 73*(2), 305–332. https://doi.org/10.1111/peps.12353

Zhu, H. (2019) kableExtra: Construct complex table with 'kable' and pipe syntax. R package version 1.1.0. https://CRAN.R-project.org/package=kableExtra