Educational and Psychological Measurement
http://epm.sagepub.com/

Modeling Socially Desirable Responding and Its Effects

Matthias Ziegler and Markus Buehner

Educational and Psychological Measurement 2009 69: 548 originally published online 15 October 2008
DOI: 10.1177/0013164408324469

The online version of this article can be found at:
http://epm.sagepub.com/content/69/4/548

Published by:
SAGE
http://www.sagepublications.com

Educational and Psychological Measurement
Volume 69 Number 4
August 2009 548-565

10.1177/0013164408324469
http://epm.sagepub.com
hosted at
http://online.sagepub.com

Modeling Socially Desirable Responding and Its Effects

Matthias Ziegler
Humboldt University Berlin
Markus Buehner
Ludwig-Maximilians-University Munich

The impact of socially desirable responding or faking on noncognitive assessments remains an issue of strong debate. One of the main reasons for the controversy is the lack of a statistical method to model such response sets. This article introduces a new way to model faking based on the assumption that faking occurs due to an interaction between person and situation. The technique combines a control group design with structural equation modeling and allows a separation of trait and faking variance. The model is introduced and tested in an example. The results confirm a causal influence of faking on means and covariance structure of a Big 5 questionnaire. Both effects can be reversed by the proposed model. Finally, a real-life criterion was implemented and predicted by both variance sources. In this example, it was the trait but not the faking variance that was predictive. Implications for research and practice are discussed.

Keywords: *social desirability; faking; validity; spurious measurement error; structural equation modeling; common method variance*

For years, cognitive assessments have been used to predict performance. Their utility to predict academic (Kuncel, Hezlett, & Ones, 2004) as well as job success (Schmidt & Hunter, 1998) has been shown. The rise of the Big 5 as a model to describe individual differences in personality self-descriptions (Goldberg, 1992) reanimated efforts to use noncognitive assessments to predict performance. By now, many studies have been conducted, demonstrating the utility of noncognitive measures in predicting academic as well as job performance (e.g., Higgins, Peterson, Pihl, & Lee, 2007). Nevertheless, noncognitive assessments are still often criticized. The main reason for the criticism is the potential influence of socially desirable responding (SDR), or faking (e.g., Griffith, Chmielowski, & Yoshita, 2007). In other words, people are afraid that the results they get from noncognitive assessments do not represent the actual characteristics of the participants. Although this issue has been investigated for many years (Crowne & Marlowe, 1964; Hogan, Barrett, & Hogan, 2007),

Authors' Note: Please address correspondence to Prof. Dr. Matthias Ziegler, Humboldt University, Psychology Institute, Unter den Linden 6, 10099 Berlin, Germany; e-mail: matthias.ziegler@psychologie.hu-berlin.de

questions regarding its impact on construct and criterion validity remain. One of the main reasons for the open questions is the problem of modeling social desirability. This article suggests a new modeling approach allowing the investigation of effects on means and covariance structure as well as criterion validity of measured traits.

As an example, we use a personality questionnaire assessing the Big 5 as one of the most important noncognitive constructs. We chose the revised NEO Personality Inventory (NEO-PI-R; Costa & McCrae, 1992) because it is a widely used and well-researched instrument. Validity-related evidence exists concerning construct (Aluja, Garcia, Garcia, & Seisdedos, 2005) and criterion validity (Piedmont & Weinstein, 1994). Moreover, it has been applied in other faking studies (Pauls & Crost, 2005b).

Effects of SDR

Faking changes the mean scores of trait questionnaires (Viswesvaran & Ones, 1999; Ziegler, Schmidt-Atzert, Bühner, & Krumm, 2007). Respondents distort answers to items in a socially desirable way. Consequently, the score of some of the items changes, and so does the overall mean. As far as the Big 5 and faking good are concerned, it is also known that not all of the factors are affected. A meta-analysis revealed that scales that are especially relevant for the job a person applies for are faked more (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). Concluding, it can be said that SDR affects the means of scales depending on the importance of each scale for the overall goal of respondents.

Faking can also affect the covariance structure of distorted scales. A study conducted by Pauls and Crost (2005b) showed simple fake good instructions increased all intercorrelations between the Big 5. A job-specific faking instruction only increased the intercorrelations among factors believed to be job relevant. Whereas the impact on mean structure seems to be without controversy, the impact on covariance structure is controversial (Ellingson, Smith, & Sackett, 2001). Two explanations can be given for the different results. First, samples containing applicants for different jobs are often combined (Hogan et al., 2007). This blurs the impact of faking on the covariance structure because it can be assumed that different traits are relevant for different jobs. Consequently, the increased trait intercorrelations differ for the specific job groups, and this makes it difficult to find increased correlations in the combined sample. A second problem is that some of the studies use social desirability scales to detect faking (Ferrando & Chico, 2001). This procedure is problematic due to the nature of such scales, which have meaningful correlations with substantial personality traits such as neuroticism, agreeableness, and conscientiousness (Ones, Viswesvaran, & Reiss, 1996; Paulhus, 1991). Hence, extreme scores in such scales do not necessarily indicate response distortion but could simply represent

extreme personalities. Moreover, correcting for faking using such scales proved to be ineffective (Ellingson, Sackett, & Hough, 1999). Even though a final conclusion might yet be impossible, there is mounting evidence that faking has an effect on the covariance structure of faked scales.

A related aspect concerns the impact of faking on criterion validity. It is assumed that faking does not influence criterion validity (Ones & Viswesvaran, 1998). Although the results seem very clear, they could be criticized for the same reasons just stated. It could again be argued that using social desirability scales to detect faking or combining different samples might blur the effects. Thus, the question of criterion validity for faking and trait variance is still not answered properly.

Summarizing, there is evidence that faking has an impact on means and mixed evidence that faking has an impact on covariance structure. At first glance, faking does not affect criterion validity. However, the impact of faking on criterion validity remains questionable as long as a clear separation of faking and trait variance is impossible.

It is also important to keep in mind that the effects of faking only apply to job-relevant scales. Consequently, faking is not just simply a response distortion. Rather, it depends on contextual variables.

Faking as a Spurious Measurement Error

Based on the observation that faking is context specific, we conclude that faking can be understood as a systematic measurement error resulting from the interaction between context (situational demand) and person. Schmidt, Le, and Ilies (2003) coined the term ''spurious measurement error'' for such interactions. Classical test theory defines a test score as the sum of true score and measurement error. The measurement error is supposed to be due to unsystematic influences. If faking is seen as a spurious measurement error, it would represent systematic variance. Spurious measurement errors are systematic because it is assumed that this error does not always occur, but it always occurs under identical circumstances. Variance due to systematic measurement errors cannot easily be distinguished from true score variance. However, whereas true score variance only increases correlations among items measuring the same trait, a spurious measurement error will increase correlations among all faked items, and thus correlations between different traits. This makes modeling spurious measurement errors and thus modeling faking possible.

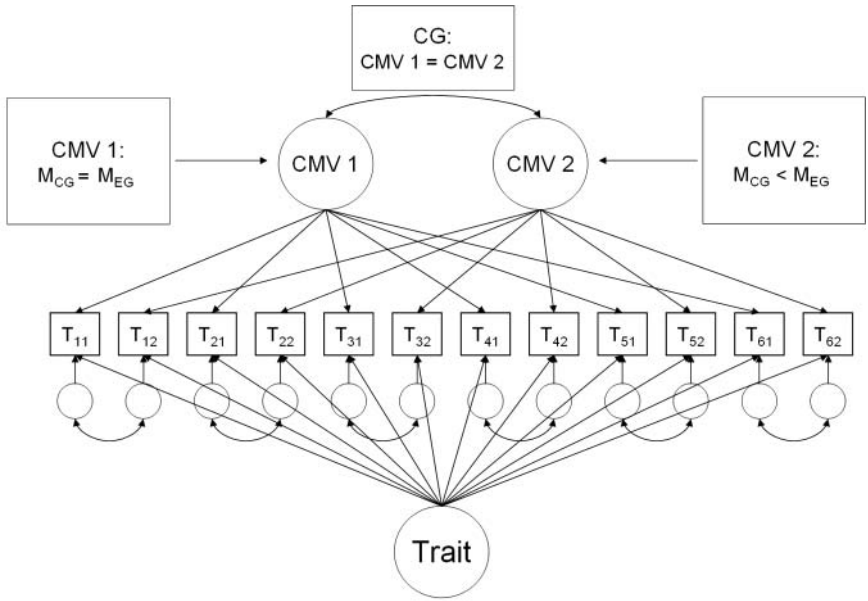
Modeling Social Desirability

We propose a new way of separating trait from faking variance. A spurious measurement error (faking) contributes to correlations between scales, however, not between scales measuring one trait but between scales faked. In this sense, a

systematic measurement error can be viewed as common method variance (CMV). Podsakoff, MacKenzie, Lee, and Podsakoff (2003) suggest modeling CMV as a latent variable using structural equation modeling. Applying such a model, all faked items or scales would not only have loadings on their specific trait variable but also on a common method factor representing faking. However, the authors cautioned users, because such a latent method factor could comprise different things. Obviously, faking can be one of the explanations. However, other possibilities, such as a higher order trait factor, common rater effects, or item characteristics, to name just a few, are also plausible. To clarify the character of the latent factor, Podsakoff et al. suggested correlating it with marker variables. For example, a common method factor representing faking should be correlated with a social desirability scale. However, this might be misleading because such scales are loaded with substantial personality differences themselves. Thus, in the case of the Big 5, a correlation does not necessarily confirm that faking was extracted. It could also mean a substantial higher order factor was extracted. We propose a different approach that allows a clear interpretation of the common method variable as faking. For this purpose, an experimental design was combined with the approach suggested by Podsakoff et al.

Two groups are administered the questionnaire twice. Spurious measurement errors should occur at both measurement points in both groups assuming that SDR always occurs to some extent. A control group (CG) is asked to respond honestly both times (low stakes), and an experimental group (EG) is given a specific faking instruction the second time (high stakes). At the first measurement point, both common method factors should have the same character. However, at Time 2, the character of the common method factor in the EG should have changed due to the specific faking instruction. This instruction should influence the interaction between context (situational demand) and person. In other words, people will fake specific scales that they believe to be important. Consequently, the intercorrelations between these scales will increase due to the higher impact of the spurious measurement error. This should also influence the loading pattern for the common method factor. If such a change occurs, it can only be due to the faking instruction, which was the only difference between both groups. To clearly interpret the factor as faking factor, though, four hypotheses must be tested. (1) The intercorrelations between faked traits should decrease controlling for faking variance. (2) Mean scores for faked traits should decrease for the same reason. (3) The mean score of the common method factor at Time 2 should be larger in the EG compared to the CG. (4) The correlation between the common method factors for Time 1 and Time 2 should be low in the EG but high in the CG. In the CG, no specific faking instruction is given; therefore, the common method factor should not change its character at Time 2. This would result in a high correlation between both factors. In the EG, however, a specific faking instruction is given, which should change the character of the common method factor resulting in a low intercorrelation. Figure 1 gives an

Figure 1
Depiction of Hypotheses



Note: CMV = common method variance; EG = experimental group; CG = control group; T_{ij} = trait indicator *i* at time point *j*.

exemplary model as well as a depiction of the hypotheses just outlined. Correlated error terms are specified because the same trait indicators are used at both measurement points. For easier understanding, we will outline the procedure based on an empirical example.

Method

Participants and Procedure

The sample consisted of $N = 341$ (270 women) psychology students. The average age and semester were 22.72 ($SD = 5.67$) and 1.89 ($SD = 2.01$), respectively. The CG contained $n = 167$ (39 men), and the EG $n = 174$ (30 men) participants. Group assignment was random. Participants first had to fill out the NEO-PI-R with regular instructions. Afterwards, a series of other tests that lasted for about 2 hr

was administered. Among those tests was the Intelligence Structure Test (Amthauer, Brocke, Liepmann, & Beauducel, 2001). Scores in this test were used to estimate missing values in the criterion variable (see below). Then participants received a specific instruction depending on their group assignment. In the CG, participants were told not to be surprised if they knew the upcoming test and to answer the questions as honestly as possible, not trying to replicate their first answers. Within the EG, participants received a specific fake-good instruction (Rogers, 1997). Thus, participants in the EG were told they were about to take part in a student-selection procedure for psychology:

> Universities have to select their students. For this task a number of instruments like the following are being tested right now. Please imagine that you are participating in a student selection procedure. Of course, it is your goal to get admitted as a psychology student. Therefore, you have to fill out the following questionnaire in a way that assures your admission. However, you have to be careful because a test expert will check the results for obvious faking and you do not want to be spotted.

After reading their specific instructions, all participants filled out the NEO-PI-R for a second time. Upon completion, they were asked whether they had followed their specific instructions, which all participants affirmed. Finally, they were thanked and dismissed. After finishing data collection, feedback and full information on the experiment was provided.

Test Materials

Both groups had to fill out the 240 NEO–PI–R items in the German version (Borkenau & Ostendorf, 1993) twice. Six facets for each of the five factors of the five factor model, that is, neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness, are assessed with eight items each. Items ask participants to rate themselves in typical behaviors or reactions on a 5-point Likert-type scale ranging from *strongly disagree* to *strongly agree*. Alphas for the factors range from .87 to .92 and from .53 to .85 for the facets.

To test the capability of the new approach to test the impact of faking on criterion validity, a criterion is needed. The faking instruction guided participants into faking in a university admission setting. Thus, the criterion should evaluate academic success in psychology. A good measure of success in psychology is the grade in statistics (Furnham & Chamorro-Premuzic, 2004), which was also used here. The examination consisted of 35 multiple-choice items ($\alpha = .91$) and took place at least 2 months after data collection. Unfortunately, not all study participants (80.1%) were in their first semester. Therefore, grades in the exam were imputed for higher semester students (19.9%), who did not take the exam. A listwise deletion would be possible because the MCar test by Little (1988) was not significant, $\chi^2(8) = 10.04$,

$p = .26$. However, to obtain maximal power, the expectation-maximization (EM) algorithm was used with intelligence test results as predictors to estimate missing values.

Models

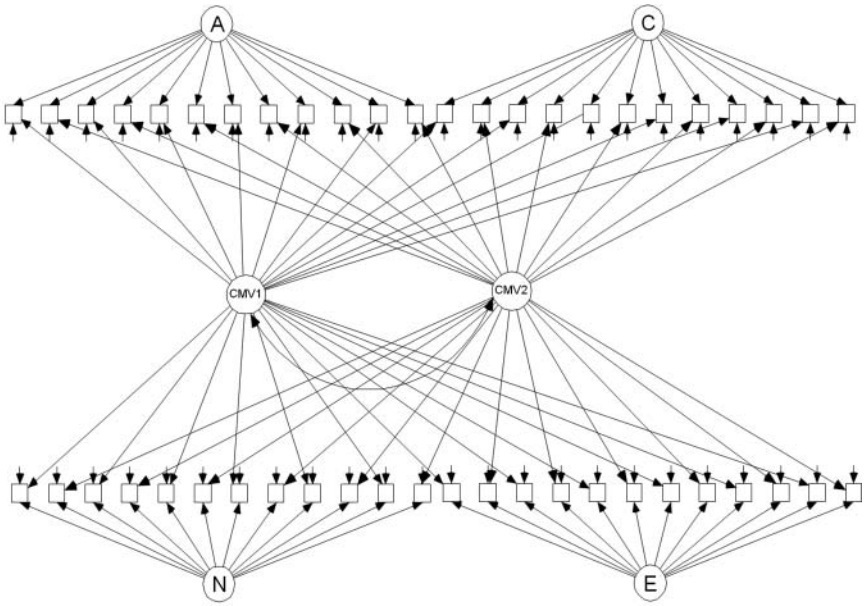
A multigroup structural equation model was specified to separate trait and faking variance. The model contained the NEO-PI-R facets (sum of eight items each) for both measurement times as indicators of their specific trait. Furthermore, two additional latent variables representing CMV1 and CMV2 were integrated, which had paths to either all indicators measured at Time 1 or all indicators measured at Time 2. Both variables were allowed to correlate. Because the common method factors at Time 1 should be equal in both groups, latent means were set equal in both groups.

To test all our hypotheses, the analyses consisted of two different structural equation models: (a) a correlated trait model and (b) a prediction model. Because it was assumed that not all personality factors are faked, the models will only contain those factors that were faked (Pauls & Crost, 2005a, 2005b). Because the error only affects some of the traits, only these traits will share faking variance. Thus, only these traits can be used to model faking. Equal items were used at both measurement times, and hence, correlated errors between identical facets were included. Facet loadings for each factor were set equal in the CG representing the finding that factor loadings under normal conditions are highly comparable (e.g., Costa & McCrae, 1992). Using the model, it will be possible to test the four hypotheses stated above. To test Hypotheses 1, correlations at Time 2 between personality factors were first computed without controlling for the spurious measurement error. They should be higher in the EG than in the CG. In a second step, the correlations were calculated modeling the common method factors. Here, the correlations should be low and comparable for both groups. All other specific hypotheses were tested with the structural equation models. Furthermore, the question of criterion validity of trait and faking variance will be explored. To do this, grades in an examination in statistics were taken into the model and regressed on the personality factors as well as the common method factor at Time 2.

Statistical Analyses

The data will be subjected to an analysis without controlling for situational demand using SPSS 14.0. To find out which of the personality factors were faked, five ANCOVAs with group as independent variable, personality score at Time 2 as dependent variable, and personality score at Time 1 as covariate were conducted. This approach proved to be the most powerful technique in designs such as this (Vickers & Altman, 2001). Even though no group differences at Time 1 were

Figure 2
Uncorrelated Structural Equation Model



Note: CMV = common method variance; N = neuroticism, E = extraversion, A = agreeableness, C = conscientiousness. Facet names of personality domains were left out. Each manifest variable represents a personality facet (sum of eight items). Each trait has 12 manifest variables, 6 for each measurement point. Correlated errors between identical items representing method variance are not depicted.

expected, as group membership was assigned randomly, this approach is more precise because it controls for even the slightest group differences at Time 1.

Confirmatory factor analyses (maximum likelihood) were conducted using AMOS 16.0. The assumption of multivariate normal distribution was violated (Mardia test: multivariate kurtosis = 129.67, critical ratio = 12.35, $p < .001$), and Bollen-Stine bootstraps with $n = 200$ samples were performed to correct the p value for the χ^2 tests.

The assessment of the global goodness-of-fit was based on recommendations by Hu and Bentler (1999) as well as Beauducel and Wittmann (2005). Thus, the standardized root mean square residual (SRMR $\leq .11$) and the root mean squared error of approximation (RMSEA $\leq .06$ for $N > 250$) were used. Because the tested models are rather complex, one should keep in mind that less restrictive cutoffs than proposed here should be chosen, as Cheung and Rensvold (2001) concluded.

Table 1
Descriptive Statistics for All Variables

Variable	Control Group *M* (*SD*)	Experimental Group *M* (*SD*)
Neuroticism T1	94.15 (26.06)	92.42 (23.16)
Neuroticism T2	90.88 (25.32)	52.37 (19.13)
Extraversion T1	117.47 (18.46)	121.73 (19.54)
Extraversion T2	116.74 (17.18)	131.84 (13.79)
Openness T1	129.63 (14.79)	131.13 (18.44)
Openness T2	128.50 (15.94)	130.43 (13.81)
Agreeableness T1	117.19 (17.19)	117.92 (17.62)
Agreeableness T2	119.04 (17.52)	129.74 (13.86)
Conscientiousness T1	119.37 (20.02)	122.98 (18.00)
Conscientiousness T2	117.81 (19.00)	153.65 (16.67)
Statistics grade	2.99 (1.39)	2.83 (1.32)

Note: T = time. Statistics grades ranged from 0.7 to 5.0, with lower values indicating better grades.

Therefore, cutoffs were applied here as general guidelines but not as strict borders. Usually, an incremental fit index is used as well. However, Beauducel and Wittmann pointed out that such fit indices tend to reject trait models and should not be used. Moreover, Cheung and Rensvold suggested not using incremental fit indices if the assumption of multivariate normal distribution is violated.

To conduct the multigroup analysis and the latent mean comparisons, the intercept for each manifest variable had to be set equal in both groups (see Byrne, 2001). All latent means were fixed at zero in the CG (except for CMV1, which was set equal in both groups).

Results

Descriptive statistics for the measures used for both groups can be found in Table 1.

Without Controlling SDR

Bivariate correlations between personality factors at both measurement points can be found in Table 2.

As can be seen, there are only small to moderate correlations between the personality factors within both groups at Time 1. The picture remained the same within the CG at Time 2. Within the EG, some of the correlations between the personality factors increased. Thus, the results show, faking led to increased correlations between

Table 2
Correlations Between Personality Factors

	N	E	O	A	C
Time 1					
N	—	–.39***	–.04	.04	–.33***
E	–.39***	—	.24*		.10
O	–.12	.44***	—	.19	–.15
A	–.19	.11	.17	—	–.01
C	–.15	.04	.06	.17	—
Time 2					
N	—	–.36***	.01	.02	–.36***
E	–.37***	—	.27**	–.01	.10
O	–.08	.54***	—	.22*	–.15
A	–.19*	.25**	.34***	—	–.01
C	–.72***	.33***	.12	.16	—
SEM[a]	N	E	A	C	
N	—	.04	–.08	–.10	
E	.03	—	–.37	.10	
A	–.06	–.27	—	.01	
C	–.08	.08		—	

Note: N = neuroticism; E = extraversion; O = openness; A = agreeableness; C = conscientiousness; SEM = structural equation modeling. Below the diagonal are the correlations within the experimental group and above within the control group for Time 1 and Time 2, respectively. Significance levels have been Bonferroni corrected for 10 tests in each diagonal.

a. Below the diagonal are the correlations within the experimental group controlling for situational demand with no correction for attenuation and above with correction for attenuation (Hancock & Mueller, 2001). Significance levels have been Bonferroni corrected for six tests in each diagonal.

*$p < .05$. **$p < .01$. ***$p < .001$.

some personality factors. This confirms the first part of the specific Hypothesis 1 and can be regarded as evidence for a spurious measurement error. It is very informative to inspect the personality trait variances in both groups (see Table 1). Within the CG, the variance remained unchanged from Time 1 to Time 2. Within the EG, however, variance actually dropped for most traits due to a ceiling effect. Usually, a restriction in variance goes along with decreased correlations. However, within the present data some correlations increased further, supporting the idea of a spurious measurement error.

The second step in the analyses was five ANCOVAs. The results showed that all personality scores except for openness, $F(1, 338) = .02$; $p = .89$; $\eta^2 < .001$; $1 - \beta = .05$, were faked. Large effects occurred for neuroticism, $F(1, 338) = 135.37$; $p < .001$; $\eta^2 = .29$; $1 - \beta = 1$, and conscientiousness, $F(1, 338) = 90.96$; $p < .001$; $\eta^2 = .21$; $1 - \beta = 1$. Small to moderate effects were observed for extraversion, $F(1, 338) = 13.86$; $p < .001$; $\eta^2 = .04$; $1 - \beta = .96$, and agreeableness, $F(1, 338) = 14.03$;

$p < .001$; $\eta^2 = .04$; $1 - \beta = .96$. Table 1 reveals that participants in the EG depicted themselves as less neurotic, more extraverted, agreeable, and conscientious than participants in the CG at Time 2 controlling for any differences at Time 1 (largest difference at Time 1 occurred in extraversion; Cohen's $d = .22$, n.s.).

Modeling SDR

To extract the spurious measurement error, openness was dropped from the following analyses. Model 1 converged properly achieving an acceptable fit: $\chi^2(2084) = 4951.52$, Bollen-Stine p value $= .01$, SRMR $= .118$, RMSEA $= .064$ (90% confidence interval [CI]: $.061 - .066$). Looking at the correlations (see Table 2) reveals only small values, mostly close to zero. In the EG, no significant correlation occurred when faking was controlled for. This confirms the second part of the specific Hypothesis 1. To compare the correlations with the bivariate correlations from Times 1 and 2, corrections for attenuation were undone (Hancock & Mueller, 2001). The only sizable correlation occurred between extraversion and agreeableness.

The model without trait intercorrelations also converged properly and achieved an acceptable model fit: $\chi^2(2096) = 4989.10$, Bollen-Stine p value $= .01$, SRMR $= .119$, RMSEA $= .064$ (90% CI: $.062 - .066$), which was significantly worse compared with the former, $\Delta\chi^2(12) = 37.58$, $p < .001$. However, model fit was not perfect. An investigation of the modification indices revealed that including correlations between some of the facets, especially extraversion and agreeableness facets, would improve model fit. This is not surprising, for two reasons. First, the correlation between extraversion and agreeableness was the only substantial one. Second, cross-loadings between facets often occur for the NEO-PI-R. The following results did not differ for the correlated and the uncorrelated model. Therefore, and because the trait intercorrelations were weak, we will report results from the uncorrelated model.

In the EG, all loadings on the personality traits were significant with few exceptions: The conscientiousness facets from Time 2, assertiveness (E3), activity (E4), impulsiveness (N5, only at Time 2), trust (A1), and positive emotions (E6, only at Time 1). Trait loadings in the CG were all significant. The differences can be explained by the different loading patterns on CMV2. Within the CG, loadings on CMV1 were significant at Times 1 and 2, except for impulsiveness (N5), excitement seeking (E5), and deliberation (C6). Within the EG, the loading pattern on CMV1 was about equal, except that insignificant loadings only occurred for order (C2), dutifulness (C3), straightforwardness (A2), compliance (A4), and modesty (A5). In both groups, loadings were small to moderate. Loadings on CMV2 were similar to loadings on CMV1 in the CG. However, within the EG all loadings were now significant with the exception of excitement seeking (E5). This supports the idea of a stronger situational impact at Time 2 in the EG. Loadings increased for neuroticism and conscientiousness facets (all $\lambda \geq |.62|$). Neuroticism facets had negative loadings on the CMV2 in the EG, whereas all other facets had positive loadings.

Table 3
Means and Variances for Latent Variables

	M		σ^2		r_{tt}	
Variable	CG	EG	CG	EG	CG	EG
N	0	−.12 (−.01)	10.99***	19.50***	.87	.80
E	0	−.28 (−.06)	3.61***	5.66***	.67	.76
A	0	.21 (.05)	6.70***	2.35*	.90	.72
C	0	.63* (.13)	7.07***	2.74***	.88	.79
CMV 1	1.16	1.16	3.81***	4.62***	.85	.88
CMV 2	0	5.71*** (.75)	9.83***	5.56***	.86	.95

Note: CG = control group; EG = experimental group; r_{tt} = reliability according to Hancock and Mueller (2001); N = neuroticism; E = extraversion; A = agreeableness; C = conscientiousness; CMV = common method variance. Means within the CG were fixed to be 0, thus, significant means in the EG represent significant group differences. An exception is the mean for State 1, which was set equal in both groups to express equal situational demand. Effect sizes according to Hancock (2001) are given in parentheses; positive values indicate a larger mean for the EG. The effect size can be compared to a Cohen's *d*. However, latent means and variances are used and interpretation guidelines are adjusted to construct reliabilities.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 3 displays means, variances, and construct reliabilities for the latent variables. It can be seen that controlling for the spurious measurement error yielded insignificant group differences in almost all personality factor means. The only significant mean differences occurred for conscientiousness. Here, a small effect was observed (Hancock, 2001). This partly confirms specific Hypothesis 2.

The difference between the means for the CMV2 factors was moderate and significant, confirming specific Hypothesis 3.

The correlation between CMV1 and CMV2 (without correction for attenuation) amounted to $r = .78$ ($p < .001$) within the CG but only to $r = .23$ ($p < .05$) within the EG. As was expected in specific Hypothesis 4, the character of the spurious measurement error factor did not change much within the CG but did within the EG.

All latent variables had significant variances at both times and in both groups. Moreover, construct reliabilities for the traits were sufficient but smaller in the EG, which is not surprising because the loadings at Time 2 were generally smaller. This would not be observed without controlling for faking, because the spurious measurement error adds to the systematic trait variance. Whereas all reliabilities for the common method factors were about equal, a larger reliability occurred in the EG at Time 2 due to the higher loadings.

All in all, the results show that SDR can be modeled as suggested. Within the model, mean differences and increased intercorrelations could be reversed. CMV2 was substantially different in the EG, evidenced by the confirmed hypotheses. Therefore, it can be interpreted as faking factor. This means that a separation of

trait and faking variance was successful, allowing a comparison of criterion validity of personality traits versus faking.

The analysis was again conducted without trait intercorrelations. Grades in the examination in statistics were regressed on all personality factors as well as CMV2 (faking in the EG). The model converged properly and the fit was acceptable, $\chi^2(2182) = 5137.90$, Bollen-Stine p value $= .01$, SRMR $= .117$, RMSEA $= .063$ (90% CI: $.061 - .065$). Five percent of the criterion variance could be explained in the EG and 22% in the CG. The difference can be explained by variance restriction for conscientiousness in the CG. Conscientiousness was one of the best predictors in both groups (EG: $\lambda = -.15$, $p = .07$; CG: $\lambda = -.28$, $p = .001$). Thus, better grades were achieved by higher conscientiousness. Besides conscientiousness, it was neuroticism that contributed to the amount of explained variance (EG: $\lambda = -.06$, n.s.; CG: $\lambda = -.35$, $p < .001$). However, CMV2 did not have an incremental validity above and beyond the personality traits (EG: $\lambda < .01$, n.s.; CG: $\lambda = -.04$, n.s.).

Thus, it can be concluded that criterion validity in both groups resulted from individual differences in personality traits but not variance due to faking.

Discussion

A combination of an experimental design with structural equation modeling was used to model faking. The approach was based on the idea that faking can be understood as a spurious measurement error with effects such as common method variance. Using two groups and two measurement points, a separation of faking and trait variance was aimed at. Results show that using the model outlined above not only reversed the effects of faking on means and covariance structure, it also achieved the wished-for separation of trait and faking variances. The modeling allows causally attributing the change in means as well as covariance structure to the influence of faking. Moreover, criterion validities for trait variance and faking variance could be compared. In line with previous findings, criterion-related validity resulted from trait variance, especially conscientiousness and neuroticism (Higgins et al., 2007), but not faking.

Modeling SDR

Within this article, we have advocated the idea that faking can be seen as a spurious measurement error that is caused by an interaction between context and person. This error increases correlations between affected trait scores. The results confirm our hypotheses and show that such an error can be statistically modeled, as we suggested. Furthermore, the model provides information on the character of

faking itself. Neuroticism and conscientiousness have the highest loadings on CMV2 in the EG, supporting the proposition by Ones et al. (1996) that social desirability is related to these traits. It could be assumed that these two traits are believed by many to be important for jobs. Hence, they would be most susceptible to faking (Konstabel, Aavik, & Allik, 2006). However, using different scenarios might change the character of the faking factor depending on what traits are seen to be important by the participants.

Implications for Personality Research

Within the Big 5 literature, there is a debate regarding the higher order factor structure (e.g., Bäckström, 2007; Digman, 1997). Whereas some research indicates a substantial trait nature, other studies conclude higher order factors might represent a general bias. The present study adds to this discussion. Using the approach suggested, trait and bias (faking) variance were separated. The present results support the notion that shared variance between Big 5 factors could be viewed as a bias. Two results underline this. First of all, controlling for faking resulted in a lack of intercorrelations between the traits. Of interest, only a small correlation between extraversion and agreeableness remained. Rost, Carstensen, and von Davier (1999) could show that extraversion is a very heterogeneous construct. Moreover, both personality traits are concerned with interactions between people. Thus, shared variance could be due to a shared theme. Second, faking variance was not predictive of the criterion as should be expected for a higher order trait. However, we only used one questionnaire as well as one specific criterion. Further research using this approach is needed before final conclusions can be drawn.

Implications for Social Desirability Research

One of the most important points of this article is the introduction of a new approach to modeling social desirability. This new approach allows distinguishing between trait and faking variance. Therefore, it is possible to closely examine the effects of social desirability on the psychometric properties of a questionnaire.

The present results seem to underline existing research indicating that faking affects construct validity but not criterion validity. However, we only used one specific questionnaire in a specific situation as well as only one criterion. It is plausible to assume that different results might occur if the questionnaire or criterion changed. For example, using supervisor ratings might reveal a substantial contribution of faking variance. To explore this, applicants can be tested when they apply as well as once they have taken a job. It should also be possible to get personality scores from job incumbents at two measurement points. The first group could be compared with our EG, whereas the second would stand as CG. One must only pay

attention to the fact that only job-relevant aspects are faked. Therefore, it is not advisable to mix different applicant samples. In this study, participants thought being open to experience is not helpful to succeed in psychology. Openness rarely is affected by faking (Birkeland et al., 2006). A possible reason for this might be that the attributes connected with openness, such as being creative, interested in many things, and full of fantasy, are not seen as essential for academic success.

Limitations

Because a new design was used, replications in new samples with different tests and criterions are needed.

A methodological limitation is given as well. To use the presented structural equation model to extract faking variance, at least two different traits must be faked by the participants. Otherwise, spurious measurement error variance and trait variance could not be separated, because trait and method factor would try to explain the same variance.

Furthermore, it could be argued that a study using students and a grade bears only little practical relevance. However, as Rogers (1997) stated in his suggestions regarding faking experiments, the sample should have practical relevance. Because the whole experimental design was customized for a university setting, this practical relevance is given. Within a workplace setting, such as an applicant setting, different results regarding factors faked and facets affected by situational demand might occur. The only drawback of the present sample is that it is partly preselected for intelligence. All psychology students have been selected by their grades in school. Thus, a more homogeneous group regarding intelligence occurred. Moreover, typically there are more female psychology students than male students.

Another critical point is model fit. The exact test shows that there is still unexplained systematic variance left. However, the fit indices show that this misfit is not large. An investigation of the modification indices suggested correlations between some of the facets that were not included in the model. Moreover, there was not a single outstanding modification index indicating one serious misspecification. Rather, there were several smaller modification indices. As explained above, we abstained from including them. This procedure actually increases chances of finding substantial trait intercorrelations. Consequently, a correlation occurred between extraversion and agreeableness—exactly those factors for whose facets the modification indices suggested correlated errors. Nevertheless, this indicates lacking divergent validity at facet level, which was already acknowledged by Costa and McCrae (1995).

Finally, we focused on faking good and not faking bad or malingering. However, the presented approach can easily be generalized to investigations of malingering.

This article introduces a new approach to modeling SDR without using specific scales or difference scores. The results support the idea that faking can be understood

as spurious measurement error. Moreover, the approach can be of use in personality as well as applied psychology. The main advantage is that a causal interpretation for the influence of faking on means, covariance structure, and criterion validity of noncognitive questionnaires becomes possible.

References

Aluja, A., Garcia, O., Garcia, L. F., & Seisdedos, N. (2005). Invariance of the "NEO-PI-R" factor structure across exploratory and confirmatory factor analyses. *Personality and Individual Differences*, *38*, 1879.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *I-S-T 2000 R (Intelligenz-Struktur-Test 2000 R)* [Intelligence-Structure-Test 2000 R]. Göttingen, Germany: Hogrefe.

Bäckström, M. (2007). Higher-order factors in a five-factor personality inventory and its relation to social desirability. *European Journal of Psychological Assessment*, *23*, 63-70.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling*, *12*(1), 41-75.

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*, 317-335.

Borkenau, P., & Ostendorf, F. (1993). *NEO-PI-R nach Costa und McCrae* [NEO-PI-R by Costa and McCrae]. Göttingen, Germany: Hogrefe.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum.

Cheung, G. W., & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods*, *4*, 236-264.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI): professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & Mccrae, R. R. (1995). Domains and facets—Hierarchical personality-assessment using the revised NEO Personality-Inventory. *Journal of Personality Assessment*, *64*, 21-50.

Crowne, D. P., & Marlowe, D. (1964). *The approval motive*. New York: John Wiley.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*, 1246-1256.

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.

Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, *86*, 122-133.

Ferrando, P. J., & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement*, *61*, 997-1012.

Furnham, A., & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and Individual Differences*, *37*, 943-955.

Goldberg, L. R. (1992). The development of marker variables for the Big-Five factor structure. *Psychological Assessment*, *4*, 26-42.

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

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373-388.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. d. Toit & D. Sörbom (Eds.), *Structural equation modeling: Present and future—Festschrift in honor of Karl Jöreskog* (pp. 195-216). Lincolnwood, IL: Scientific Software International.

Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology*, *93*, 298-319.

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

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

Konstabel, K., Aavik, T., & Allik, J. (2006). Social desirability and consensual validity of personality traits. *European Journal of Personality*, *20*, 549-566.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, *86*, 148-161.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*, 1198-1202.

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

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*, 660-679.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson & L. S. Wrightman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.

Pauls, C. A., & Crost, N. W. (2005a). Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. *Journal of Individual Differences*, *26*, 194-206.

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

Piedmont, R. L., & Weinstein, H. P. (1994). Predicting supervisor ratings of job performance using the NEO Personality Inventory. *Journal of Psychology: Interdisciplinary and Applied*, *128*, 255-265.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879-903.

Rogers, R. (1997). Researching dissimulation. In R. Rogers (Ed.), *Clinical assessment of malingering and deception* (2nd ed., pp. 309-327). New York: Guilford.

Rost J., Carstensen, C. H., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R.E: Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 324-332). New York: Waxmann.

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*, 262-274.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual-differences constructs. *Psychological Methods*, *8*, 206-224.

Vickers, A. J., & Altman, D. G. (2001). Statistics notes—Analysing controlled trials with baseline and follow up measurements. *British Medical Journal*, *323*, 1123-1124.

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

Ziegler, M., Schmidt-Atzert, L., Bühner, M., & Krumm, S. (2007). Fakability of different measurement methods for achievement motivation: Questionnaire, semi-projective, and objective. *Psychology Science*, *49*, 291-307.