Scandinavian Journal of Psychology, 2015, 56, 582–591 DOI: 10.1111/sjop.12231

Personality and Social Psychology

Correction for faking in self-report personality tests

LENNART SJÖBERG

Center for Media and Economic Psychology and Center for Risk Research, Stockholm School of Economics, Sweden

Sjöberg, L. (2015). Correction for faking in self-report personality tests. *Scandinavian Journal of Psychology, 56*, 582–591.

Faking is a common problem in testing with self-report personality tests, especially in high-stakes situations. A possible way to correct for it is statistical control on the basis of social desirability scales. Two such scales were developed and applied in the present paper. It was stressed that the statistical models of faking need to be adapted to different properties of the personality scales, since such scales correlate with faking to different extents. In four empirical studies of self-report personality tests, correction for faking was investigated. One of the studies was experimental, and asked participants to fake or to be honest. In the other studies, job or school applicants were investigated. It was found that the approach to correct for effects of faking in self-report personality tests advocated in the paper removed a large share of the effects, about 90%. It was found in one study that faking varied as a function of degree of how important the consequences of test results could be expected to be, more high-stakes situations being associated with more faking. The latter finding is incompatible with the claim that social desirability scales measure a general personality trait. It is concluded that faking can be measured and that correction for faking, based on such measures, can be expected to remove about 90% of its effects.

Key words: Self-report personality test, faking, correction for faking, high-stakes testing.

Lennart Sjöberg, Center for Media and Economic Psychology and Center for Risk Research, Stockholm School of Economics, Box 6501, 113 83 Stockholm, Sweden. Tel: +4687557234; e-mail: lennartsjoberg@gmail.com

INTRODUCTION

Most personality tests used in the workplace are of the self-report type. The test person's task is to decide whether an assertion formulated as an item applies or does not apply to him or her. Rarely, if ever, is there any possibility to check if the answers are truthful, and there are strong reasons to believe that many test takers are enhancing the image they give of themselves, especially in high-stakes situations. There is general agreement that faking occurs and that it is strong in high-stakes situations, probably also in interviews (Roulin, Bangerter & Levashina, 2014). Explicit Big Five scales are no exception, although implicit associations measures may be less vulnerable to faking (Vecchione, Dentale, Alessandri & Barbaranelli, 2014). Faking is an important problem in the application of self-report personality tests.

The issues of faking involve three kinds of questions:

1. What is its prevalence?
2. What are its consequences?
3. What can be done about it?

OVERVIEW OF THE ISSUES

Correction for faking has often been discussed in the research literature. Two points were raised in a debate on personality tests published by Morgeson, Campion, Dipboye, Hollenbeck, Murphy and Schmitt (2007). They suggested that correction for social desirability responses is not effective because all test takers fake to the same extent. However, data show that there is considerable variation between individuals in how much they fake. Furthermore, Morgeson *et al*. suggested that faking may have a positive predictive value and therefore one should not correct for its effects, but rather seek to use it as a prognostic variable, but they presented no substantial empirical support for the claim. Knowingly providing false information about oneself on a personality test is dishonest and could probably be used as an inverse measure of honesty and integrity. Indeed, Donovan, Dwight and Schneider (2014) found that fakers had a lower level of job performance.

The consequence of faking is that some of the tested persons, those that enhance the image of themselves, gain advantages over those that do not. In preliminary research it was found that women and immigrants tend to belong to the latter group (Sjöberg, 2010), so faking may be a danger to equality and diversity in the workplace to the extent that personality tests are used in the recruitment process. Tests that lack protection against faking often provide very high scores in high-stakes situations, so high that they differentiate very little among test takers. Some test constructors try to protect themselves against such an outcome by using an ipsative response format (comparative responses) (Stark, Chernyshenko, Drasgow *et al*., 2014), but research shows that much of the impact of faking remains, and that such tests take longer to respond to and are disliked by the test persons. Ipsative formats are associated with psychometric and statistical problems (Meade, 2004). In addition there is some evidence that the results on ipsative tests are correlated with intellectual ability (Matthews & Oddy, 1997), which is undesirable since the total predictive power of personality and ability testing is higher when the two are uncorrelated than when they are correlated, all other conditions equal.

One method to deal with faking is to warn that it can be detected and that detection may have negative consequences for the tested person. However, research has not yielded unequivocally positive results for that method (Fan, Gao, Carroll *et al*., 2012) and relatively little research has been reported on the use of warnings in high-stakes situations where faking is especially common and strong.

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Response time for each test item may be an indicator of faking; the longer the response time, the more likely is it that the test taker fakes his or her responses. This assumption has to some extent been verified in research (Fine & Pirak, 2015), but the effect is weak and probably not practically useful as a basis for correction for faking (Holden & Hibbs, 1995).

The response scale is another potentially important factor. The use of only two response categories may increase faking, as compared with a Likert scale with several steps (Khorramdel, 2014). Objective personality tests (OPTs) may turn out to be less affected by faking (Ortner & Schmitt, 2014) than self-report tests. However, OPTs are very diverse and validity evidence is relatively scarce.

In the present paper, use is made of social desirability scales to correct for faking. Some researchers have argued that scales of social desirability should not be used for correction for faking because they may personality scales (Ziegler, Maccann & Roberts, 2012). However, this is probably a misleading argument. For example, it was found in one study (Study 4 below) that amount of faking varied regularly as a function of how important the test situation was to the test takers, a finding that is not consistent with a personality interpretation of this variable. See the General Discussion for further comments on this type of critique of social desirability as a measure of faking.

Correction for faking should increase the validity of a test, but this is rarely true (Ones, Viswesvaran & Reiss, 1996; Schmitt & Oswald, 2006). Validity in the sense of the correlation between a test and a criterion, however, is a crude measure. The important thing is that the uncorrected test results tend to give a group of tested persons at the top of the distribution that is to a large extent made up of fakers, which is usually undesirable when the test is used in a selection process. It is sometimes argued that faking shows that the tested person is motivated for the job or aware of what its demands are and that such motivation and awareness have positive prediction value. However, it has not been shown that faking is a positive predictor of job performance. The opposite may be true in some cases (Donovan *et al*., 2014).

MEASUREMENT OF FAKING

In the following, an overview of research is provided on the use of social desirability scales in the correction for faking on self-report personality tests. Some results on the properties of the social desirability, or faking, scales are first reported.

The basic idea for measuring faking used in the present paper is that of social desirability, as exemplified by the classical Crowne–Marlowe scale (Crowne & Marlowe, 1960). In the studies reported below, a scale constructed on the same principles, here called Overt Faking, was used. However, it can be suspected that some sophisticated test takers understand that some items belong to a social desirability scale. For that reason, a scale of ordinary personality items was constructed, selected from a large pool of such items, which correlated strongly with the scale of Overt Faking but did not have a content which could be easily identified as a faking measure by someone with knowledge of test theory. This scale is termed Covert Faking. The correlations among the Crowne–Marlowe scale, Overt Faking and Covert Faking in a group of 159 test takers are given in Table 1.

Table 1. *Correlations among three faking scales (N = 159). Job applicants tested on the Internet (Sjöberg, 2014)*

	Crowne–Marlowe's social desirability scale	Overt Faking scale	Covert Faking scale
Crowne–Marlowe's social desirability scale	1.00	0.76	0.73
Overt Faking scale	0.76	1.00	0.56
Covert Faking scale	0.73	0.56	1.00

The table shows high correlations among the three faking scales. Hence, the Overt and Covert faking scales were successfully construct validated with regard to the Crowne–Marlowe scale and they are both included in a personality test, *UPP* (Bergman, Sjöberg, Lornudd & von Thile Schwartz, 2014), which is used in studies 2–4 below. It is important to note that personality scales in the test were differentially correlated with both faking scales, and that the two sets of correlations were strongly related, see Fig. 1, which is based on data from 296 job applicants (Study 2).

Social desirability scales have often been used to measure faking. The present approach is unusual in using also a covert scale, and in applying a statistical model for estimating test scale value where social desirability variance has been removed. One drawback of the present approach to dealing with faking is the need to include separate scales. Is it possible to measure faking without separate scales? To investigate this possibility, all items of the *UPP* test, except the two faking scales, were divided into those expressing positive behavior (101 items) and those expressing negative behavior (103 items), in data from 423 job applicants. Endorsement of positive items and rejection of negative items, regardless of other aspects of their contents, can be expected to be an expression of faking. Mean responses to positive and negative items were therefore computed to form indices.[1] The reliabilities of these two indices were 0.91 and 0.93, respectively. Their correlation and correlations with Overt and Covert Faking scales are given in Table 3. The table shows

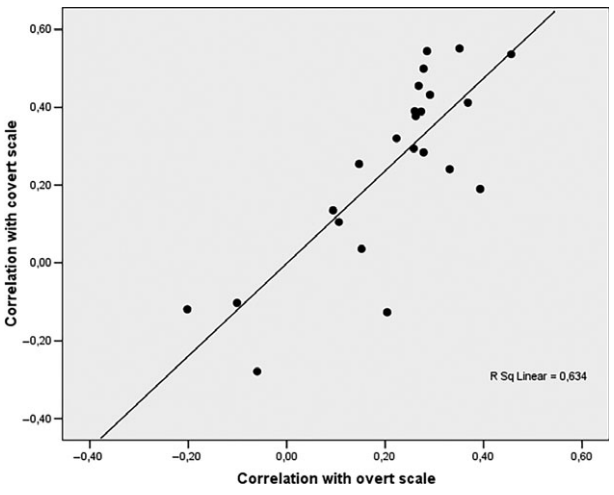

Fig. 1. Correlations between Overt and Covert faking for the 23 scales of the *UPP*[T]test.

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

substantial overlap between the variables, suggesting that responses to positive and negative items can be used as measures of faking.

The multiple correlation between the two item indices and faking scales were 0.54 and 0.74 for Overt and Covert faking, respectively. These results suggest that it should be possible to correct for faking on the basis of over-all endorsement/rejection of items. As an example of faking correction based on the two item indices, Emotional Stability was corrected for faking based on faking scales, and on item indices.[2] The two sets of residuals correlated 0.80. Clearly, these two different approaches to correction for faking gave similar results. These results can be compared to scoring "blatant extreme responses," a related approach for measuring faking, which seems to have some promise (Levashina, Weekley, Roulin & Hauck, 2014).

Data will in the following be corrected for faking on the basis of a regression model, using the two faking scales as independent variables, see Appendix for details. Four empirical studies of faking, intended to further investigate the validity of the approach used to correct for faking, will now be presented.

STUDY 1: ANALYSIS OF FAKING IN HIGH-STAKES ADMISSION TESTING

The purpose of this initial study was to investigate faking in a high-stakes testing situation, its extent and whether its effects could be eliminated by the proposed method. The test takers were either applying for admission to the Stockholm School of Economics, or they were incumbents who had already been admitted to the School. For the first mentioned group, the consequences of testing were potentially very important since they would have an impact on their chances to be admitted to the School, a very desirable consequence for most of them, for the second group none at all. For them, testing was completely anonymous; they knew that they would not even themselves be allowed to see their results.

METHOD

Measured constructs

Faking was measured with three social desirability scales:

1. Crowne–Marlowe social desirability scale (Crowne & Marlowe, 1960);
2. Paulhus Faking scale (Paulhus, 1991); and
3. Paulhus Self Deception scale (Paulhus, 1991)

The validity of the Paulhus Faking scale for measuring faking was supported by (Miller & Ruggs, 2014).

The following personality scales were used:

1. Schutte *et al*. EQ (Schutte, Malouff, Hall *et al*., 1998);
2. Alexithymia (Bagby, Parker & Taylor, 1994; Weinryb, Gustavsson, Åsberg & Rössel, 1992);
3. Self actualization (Jones & Crandall, 1986);
4. Machiavellianism (Christie & Geis, 1970); and
5. Empathy (Mehrabian & Epstein, 1970)

Big Five scales (Wiggins & Trapnell, 1997) were also included: Agreeableness; Emotional stability; Extraversion/introversion; Openness; and Conscientiousness.

Participants

One hundred and ninety participants had taken the tests as part of a process for assessing applicants to the Stockholm School of Economics. They had been invited to take the test, mainly on the basis of high school grades or a test of intellectual ability. Even if instructions stressed that they should give honest and frank answers to self-report items, it was not expected that most of them would be entirely open and frank. Of course, it was also expected that there would be variability in the extent of faking.

The group of applicants consisted of 102 men and 88 women, average age 20.5 years (range 18–34). They were comparable to those who had previously been admitted to the school, only slightly lower in grades or results on tests of intellectual ability. The latter circumstance caused no concern about lack of comparability because the kinds of personality variables studied do not correlate strongly, if at all, with academic intelligence in the traditional sense of the word.

Forty-one participants were recruited among students at the Stockholm School of Economics. They were on the average 21.1 years old (range 18–28); 19 were women, 21 men. One participant did not state gender.

Procedure

Testing was done in one session, with all participants present at the same time. The incumbent test takers were paid SEK 400 for participation (at that time about US$40).

RESULTS

Did the respondents of the testing sessions, which could have very real consequences (called High-Stakes Testing in the following), differ from those who were tested anonymously? Consider first the three scales used to measure faking, see Table 2.

These results are encouraging because they show that the faking scales all worked as expected, exhibiting very large (about 1 standard deviation) and statistically significant differences between high-stakes and low-stakes testing.

The next question is to what extent the various personality measures were affected by faking; see Table 3, which also shows the results of correcting the differences for faking.

Table 2. *Faking scores in two groups, all measures standardized to mean = 0 and SD = 1 in the combined group*

Scale used to measure faking	Mean, high-stakes testing	Mean, anonymous testing	t	df	p
Crowne–Marlowe social desirability	0.20	−0.93	7.29	229	< 0.0005
Paulhus Faking	0.15	−0.70	5.23	229	< 0.0005
Paulhus Self deception	0.15	−0.68	5.04	229	< 0.0005

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

Table 3. *Test scores in two groups, all measures standardized to mean = 0 and SD = 1 in the combined group*

Test variable	Mean, high-stakes testing	Mean, anonymous testing	*t*	df	*p*	Adjusted difference	*t* of adjusted difference
Schutte *et al*. EQ	0.16	−0.73	5.43	229	< 0.0005	0.02	ns
Alexithymia	−0.17	0.80	6.09	229	< 0.0005	0.07	ns
Self actualization	0.18	−0.82	6.32	229	< 0.0005	−0.05	ns
Machiavellianism	−0.14	0.67	4.96	229	< 0.0005	0.12	ns
Empathy	0.00	−0.02	0.00	229	ns	0.07	ns
Agreeableness	0.13	−0.62	4.55	229	< 0.0005	−0.17	ns
Emotional stability	0.19	−0.86	6.65	229	< 0.0005	−0.03	ns
Extraversion/introversion	0.15	−0.71	5.30	229	< 0.0005	0.13	ns
Openness	0.21	−0.99	7.90	229	< 0.0005	−0.18	ns
Conscientiousness	0.18	−0.82	6.24	229	< 0.0005	−0.47	2.78**

Notes: Difference between mean residuals when the four fakings and faking variables have been controlled for with regression models of faking (see Appendix). **$p < 0.01$.

All the differences in Table 3, with the exception of Empathy, are very large. The mean difference before correcting for faking was 0.50, after correction it was 0.05. Thus, only 10% of the effect of faking remained. This is a result which agrees well with the fact that the two groups also differed – even more strongly – on measures of faking and self-deception. In other words, statistical control was sufficiently strong to remove the motivational effects of the high-stakes testing situation. The only case where this was not true was that of conscientiousness. However, even in that case about half of the effect of the high-stakes situation was removed. The reason for the relative failure of this particular variable, as distinguished from all others tested for the influence of faking, reflects the fact that the measure of faking did not have any strong effect on it, contrary to the large effects found on other variables.

CONCLUSION

It is concluded that (a) faking was very strong among applicants, and (b) a very large share of the effect of faking on test results was eliminated by the use of scales of social desirability and regression models.

STUDY 2: EXPERIMENTALLY INDUCED FAKING

Study 1 supported the notion that faking could be measured. Faking scales could be used to eliminate about 90% of its effects on all investigated personality scales, with one exception. However, the data were the results of a "natural experiment," that is, there was no random assignment of subjects to the two conditions. The conclusions would be strengthened if similar results would be obtained in a strictly experimental study with random assignment to treatment and control groups. The present study reports the results of using such an approach with the *UPP* test. This test includes a number of personality scales and also scales measuring work related attitudes, such as work motivation and job satisfaction.

In the present experimental study a number of people, paid a fee for participation, were invited to take the *UPP* test either with the usual instructions to answer honestly or with an instruction in which they were asked to "fake good." The two groups were based on random assignment. The instructions to the "fake good" group were as follows:

Think about a job that you would very much like to have. Now imagine that you applied for that job and that this testing is a very important part of the hiring procedure. Answer the test items so that you appear to be a person who is exactly the kind they are looking for. It might mean that you fake a lot, but that's exactly what we intend to study in this investigation. Feel therefore free to respond tactically!

Two hypotheses were formulated:

Hypothesis 1. The two groups will differ in raw scores so that those who faked good will gave more socially desirable answers. Extensive previous research supports this hypothesis (Stanush, 1997). The tendency is expected be evident in both faking scales of the *UPP* test (Overt and Covert).

Hypothesis 2. After correcting for faking no or only small differences will remain between the groups.

METHOD

Participants

The number of participants in this study was 133, 63 men and 70 women. The median age was 23, variation 18–54 years. Their levels of education were:

- Grammar school: 9;
- High school: 89;
- College less than 3 years: 14; and
- College 3 years or more: 21.

RESULTS

Tables 4 and 5 show the average values in the test variables under the two different instructions. The two faking scales correlated 0.59. The table shows large and highly significant ($p < 0.001$) differences between the groups in faking. Before correction there were significant differences between the two groups in 7 of the 13 personality variables, after correction

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

Table 4. *Mean uncorrected and corrected scale values. Experimental data. Scales were standardized to mean = 0, standard deviation = 1*

	Uncorrected scale values		Scale values corrected for faking	
	Honest answers	Faked answers	Honest answers	Faked answers
Extraversion	−0.19	0.14	−0.02	0.02
Agreeableness	−0.28	0.21	0.00	0.00
Emotional stability	−0.24	0.18	0.02	−0.02
Openness	−0.14	0.11	0.02	−0.01
Conscientiousness	−0.11	0.08	0.14	−0.11
Endurance	−0.27	0.20	0.06	−0.04
Willingness to cooperate	−0.18	0.13	0.03	−0.03
Positive basic attitude	−0.20	0.15	−0.20	0.15
Self confidence	−0.33	0.25	−0.11	0.08
Social ability	−0.33	0.24	−0.11	0.08
Emotional intelligence	−0.19	0.14	0.04	−0.03
Creativity	−0.23	0.17	−0.02	0.01
Perfectionism	−0.13	0.09	−0.10	0.08
Mean	−0.22	0.16	−0.02	0.01
Overt faking scale	−0.30	0.22		
Covert faking scale	−0.40	0.30		

Table 5. *Conditions of testing for five groups, Study 4*

Group	Number of tested persons	The test results were expected to influence the admission decision	It was clearly stated that testing was voluntary	Testing was performed within the admission program	Testing was anonymous	Testing was administered by the Defence Force	Degree of stakes, and expected level of faking and involvement
Norm	1269	No	Yes	No	Yes	No	Very low
Incumbents	160	No	Yes	No	No	Yes	Low
Applicants 2010	56	Possibly	Yes	No	No	Yes	Rather low
Applicants 2012	130	Possibly	Yes	Yes	No	Yes	Rather high
Applicants 2011	218	Yes	No	Yes	No	Yes	High

none. The correction eliminated 93% of the effect of faking, which before correction averaged 0.38 standard deviation units. The results for work related attitudes were similar. Before correction 3 of the 6 work-related variables showed significant differences between groups, after correction none. The correction eliminated 86% of the effect of faking. Figure 2 summarizes the results.

The effects of the correction are seen in Fig. 2 for both personality scales and work-related scales. The figure shows that virtually all of the differences between the two groups were eliminated by correction.

CONCLUSIONS

Both hypotheses were confirmed. There were significant effects of the instruction to provide tactical responses in the test scales, and the two measures of faking were able to capture and eliminate almost all of these effects. The correction managed to eliminate about 90% of the effects of faking.

It is possible that the instruction to respond tactically creates other kinds of responses than those that are chosen in a high-stakes situation. However, the results are fully comparable to the ones obtained in a real field situation in Study 1 of a student admission case. Further work on such situations, using the Overt and Covert faking scales, will be reported in the following.

STUDY 3: A FIELD STUDY OF JOB APPLICANTS

It was desirable to investigate if the findings of previous studies could be replicated in a large-scale field study of job applicants. Data were provided by a major recruitment company in Sweden, which had been using the *UPP* test routinely for several years.

METHOD

Participants

Data from applicants for jobs as managers (both public and private sectors) are analyzed here and compared to data collected with the same test but under low-stakes conditions. There were 127 persons who had taken the test under low-stakes conditions, 50.4% women, age 23–63. The job applicants were 296 persons, 57.7% women, age 23–63. Their levels of education were similar, only somewhat higher in the group of applicants. In the low-stakes condition, 58.5% had a college education, as compared to 69.3% of the job applicants.

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

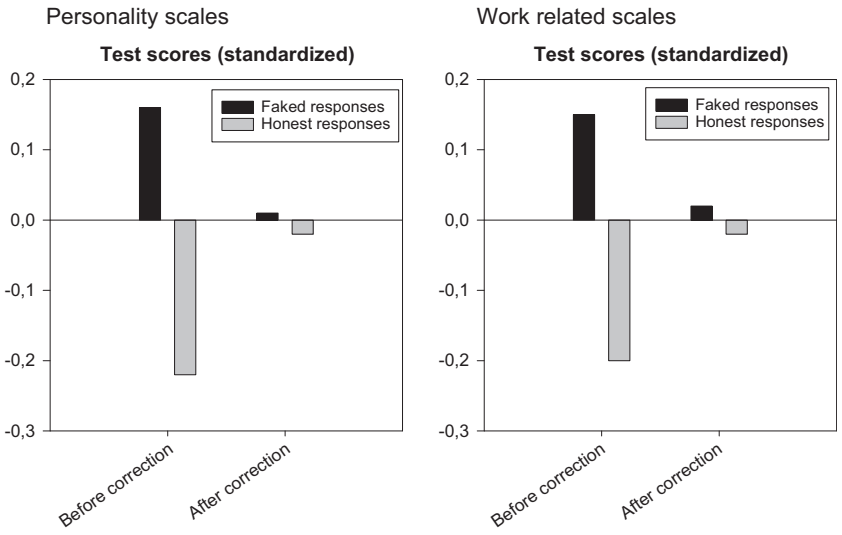

Fig. 2. Average effects of correction in Study 2, standardized scales.

RESULTS

Both Overt and Covert faking differed as expected between the low-stakes and high-stakes groups. The sizes of the differences were substantial and they were highly significant. The 24 scales of the *UPP* test were transformed according to the employed regression model (see Appendix), and means were computed for low and high-stakes conditions. In Fig. 3 there is a plot of differences between high and low stakes conditions. The figure shows that the differences were reduced quite strongly in most cases. The average difference high–low stakes was 0.22 before correction for faking, 0.04 after. In other words, about 82% of the effects of consequences of testing were eliminated by the correction procedure.

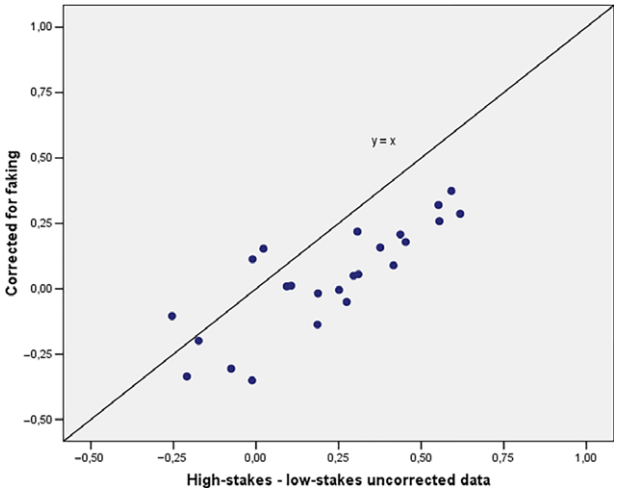

Fig. 3. Mean differences between high-stakes and low-stakes test data, corrected and uncorrected for faking.

STUDY 4: APPLICANTS FOR OFFICER TRAINING: DEGREE OF INVOLVEMENT AND FAKING

There are degrees of how important testing is perceived to be by those who take the test. High-stakes testing implies that the test results are expected to have important consequences, for example, for the chance to be accepted to training or to find employment, but just how high the stakes are, is probably varying. Voluntary or anonymous testing would seem to good examples of low-stakes situations. Another example is that of testing incumbents. The actual context, taking a test at the request of the employer, may be likely have an effect on level of faking, implying a certain degree of involvement, but weaker than if it is clear that the test results may be of personal importance.

METHOD

In the present study the degree of faking of groups that took the test in situations with varying degrees of stakes is compared. Specifically, five groups of applicants (or incumbents) to officer training in the Swedish Army, which took the *UPP* test under various conditions are compared (Sj öberg & Wolgers, 2012), see Table 6. Different degrees of faking, depending on the various conditions of testing were expected. Involvement was measured with a mood scale (Sjöberg, Svensson & Persson, 1979). The scale measures three basic mood dimensions: calmtense (tension), tired-active (activation) and sad-happy (hedonic tone). Table 5 describes the circumstances and characteristics of five different test groups, ordered in increasing degree of stakes. Table 6 gives the means of the faking scales and the mood measure of level of activation.

The table shows that level of activation and faking varied as hypothesized, as a function of the level of importance of the test results. Table 7 gives the values of eta squared for each of the Big Five scales and the variation among groups, before and after correction for faking. (The results were similar for other scales). These values exhibit the proportion of variance accounted for by groups. The table shows that variation among groups after correction was very small, but sizable before correction.

Only 11% of the group differences remained after correction for faking. As an example, mean scores in emotional stability for the five groups are given in Fig. 4. Very large effects due to varying degrees of stakes and involvement were successfully eliminated with the method used here, similar to the results of Studies 1-3.

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

Table 6. *Means of faking and activation, all scales standardized to mean = 0 and standard deviation = 1*

Group	Level of activation	Overt faking	Covert faking
Norm	−0.03	−0.24	−0.29
Incumbents	−0.42	−0.13	0.04
Applicants 2010	−0.11	0.32	0.31
Applicants 2012	0.25	0.63	0.63
Applicants 2011	0.24	0.83	0.91
One-way Anova test of group differences	$F(4,896) = 13.179$, $p < 0.0005$, eta = 0.056	$F(4,1638) = 81.969$, $p < 0.0005$, eta = 0.167	$F(4,1638) = 114.756$, $p < 0.0005$, eta = 0.219

Table 7. *Proportion of variance accounted for by groups before and after correction for faking*

Scale	Before correction	After correction
Agreeableness	0.122	0.007
Openness	0.042	0.027
Emotional stability	0.147	0.003
Extraversion	0.133	0.006
Conscientiousness	0.147	0.024
Mean	0.118	0.013

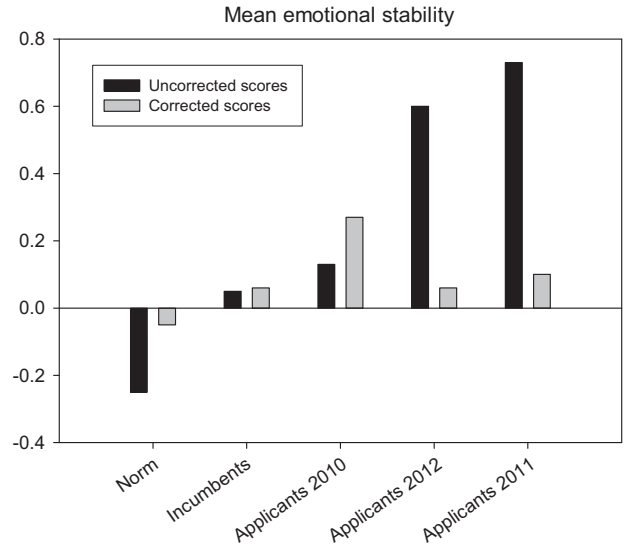

Fig. 4. Mean emotional stability for five groups, corrected and uncorrected for faking.

GENERAL DISCUSSION

The debate about faking

Discussion about faking on self-report personality tests is extensive in the literature. The following groups of arguments can be distinguished.

Faking is common and strong, can not be corrected for, and renders the tests useless. This is probably a fairly common notion. It is easy to realize that test takers can answer dishonestly, and that nothing prevents some of them from doing it if the stakes are high. Against this it may be stated that there is research, for example, as summarized in the present paper, which provides support for a more nuanced view and has proved that it is possible to overcome the problem – maybe not entirely, but largely.

Faking does not occur, or only very rarely. This is a view that is based on a few research studies that seem to have shown that applicants or incumbents responded to a test in very similar ways, for example, Hogan, Barrett and Hogan (2007). These results are exceptional, because extensive research has shown that applicants and incumbents differ strongly, not only in personality tests but probably in all contexts perceived to be value relevant, such as reported health status (Bäccman, Sjöberg & Almqvist, in press).

Faking sometimes occurs but is irrelevant to the practical value of the test. This claim is based on research that seems to have shown that the tests do not have lower validity if faking occurs, or that correction for faking does not increase validity (Ones *et al*., 1996). However, the argument ignores that very large effects often result from faking in individual cases. Top scorers on the test have often faked their responses. In addition, overall correlation between a test and a criterion of work performance is an insensitive measure of the effects of faking.

Faking occurs, but the conclusions of the tests should be based on "profiles" and they cannot be distorted by faking. However, "Profiles" may well be affected by faking. The effects of faking are not general but vary across scales.

Faking can be effectively countered by ipsative response formats. Ipsative formats do not counteract faking effectively and have a number of disadvantages: test takers dislike them, and personality results on ipsative tests can be contaminated with intelligence.

Faking can be effectively countered by warnings. Warnings have to some extent the intended effects of reducing faking (see e.g., Kovačić, Galić & Andreis, 2014), but the warnings failed to increase validity in one study (Robson, Jones & Abraham, 2008), and test takers may have a negative attitude towards them (Converse, Oswald, Imus, Hedricks, Roy & Butera, 2008).

Scales that measure the "social desirability" are in fact measurements of "personality". This argument is untenable for two reasons. First, the personality dimension usually referred to is

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

"need for approval," but independent measures of this dimension have been found to be uncorrelated with the traditional Marlowe–Crowne scale for measuring social desirability (Barger, 2012). Second, there is weak logic in the argument that a scale is invalidated if it correlates with something else that was not intended or foreseen. Suppose a scale for measuring faking happens to correlate with body weight or intelligence. Is that evidence that the scale has failed in its purpose? It is simply irrelevant to the question whether the scale measures what it is intended to measure.

Faking should be measured as the difference between the two test sessions, one with the instruction to fake, the other with the normal instruction to answer honestly. The validity of this measure can be questioned. Faking when directly asked to do so can be different from doing so spontaneously in a high-stakes situation. The difference measure cannot be expected to correlate with a third variable (in this case, the SD scale) if data from the two test sessions correlate to the same extent with the third variable. Even more important, the difference measure in a repeated-measures design should be expected to have a very low reliability due to high correlation between the two repeated measures. Hence, it cannot be expected that faking measured as a difference variable in a repeated-measures design should be correlated with an independent scale of faking, such as social desirability.

Even if faking occurs, we cannot know about it or measure it because we do not know the "true" values of the personality; nor can we correct for faking. This is an expression of an epistemological pessimism that might be defended on philosophical grounds, but it does not lead to a constructive insight or to problem solving in an applied situation. All non-physiological constructs in psychology refer to something which cannot be observed directly; they can still be investigated and knowledge about them can be acquired. Faking is not a concept with a unique epistemological status.

DISCUSSION AND CONCLUSIONS

The methodology to remove the effects of faking as advocated in this paper is based on the following notions:

1. Correction for the faking is done separately for each test scale with models based on empirical data. Different scales require different models for correction.
2. Faking is measured not only in the traditional way with a scale of "social desirability" but also with a Covert scale that can not easily be spotted by the test persons. The two scales have yielded very similar results.

Validation of this approach has been reported in the present paper. Correction eliminated about 90% of the impact of the consequences of testing, which in turn was shown to have a very large effect on the degree faking. A high-stakes condition gave rise to much more faking than low-stakes; different degrees of stakes had varying strength of effects.

A credible and validated solution of the problem with faking is very desirable. In fact, many test providers use some form of scales that measure faking (Goffin & Christiansen, 2003), but in most cases it seems that no effective method is used to exploit the information they give. If nothing is done about faking, self-report personality tests will probably increasingly be perceived as arbitrary and provide a way for people to start a career by bluffing. This is true even if one can refer to research that suggests that the validity of the test as a whole is not affected by faking.

A book edited by Ziegler *et al*. (2012), contains a number of interesting proposals for new methods to detect whether subjects fake, but are all still at a too early stage of to be practically useful; nor has any of them been proved to work better than the traditional measures of faking. A few examples:

(1) Paulhus writes about "overclaiming" (Paulhus, 2012), a term referring to the use of questions to a person if he or she is aware of a number of listed concepts, such as writers. Some of the concepts in the list are "teasers," they do not exist in reality. If you still claim to know them, that may be a sign that you are bluffing. There is some support for this technique but it has not been shown to be practically useful in a testing context.
(2) A chapter by Zickar and Sliter (2012) provides an overview of the attempts to measure faking by formulating and testing models of testing behavior ("item response models") (Zickar & Sliter, 2012). The idea behind this approach is that test takers whose data do not fit well into the model may be faking. It appears that the success of this approach is very limited; it is not practicable and not better than traditional methods.
(3) The use of response time measurements might be a way to reduce the effects of faking, but is not a practical methodology and the effects appear to be weak.
(4) Assessments of personality by other people who know the assessed persons may be less affected by faking, but is hardly a practical solution.
(5) A fairly common way to manage faking seems to be to exclude those subjects with the highest value on such a scale, such as the 25% highest. This method is weak and has insufficient power, see Reeder and Ryan (2012). Much of the effect of faking remains.

SUMMING UP

Does faking occur, is it strong enough to be of practical significance in high-stakes testing? The answer is an unequivocal yes. Does faking have effects on decision making on the basis of testing, for example in selection? The answer is yes. Does faking have importance for norm data? The answer is yes. If the test is to be used in real-life situations norm data should be collected in similar real-life situations, or they should be corrected for faking. The effect of faking seems often to be ignored; norms are collected in low-stakes situations.

It has been shown in the present paper that there is methodology to detect and correct for the faking that is sufficiently promising to live up to the demands that it should be theoretically and empirically well-founded, and practically useful. Correction with scales that measure faking works, as demonstrated

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

here, although even this method does not completely eliminate the effects.

The following persons contributed to some of the studies reported here: Elisabeth Engelberg (Study 1), Kristiina Möller (Study 2), Conny Besterman (Study 3) and Gerhard Wolgers (Study 4). Henrik Nilheim did all the necessary IT work. Financial support was provided by the Stockholm School of Economics (Study 1).

NOTES

[1] The two indices based on positive and negative responses did not include the items measuring Emotional Stability, nor did they include items measuring Overt and Covert Faking.

[2] These test takers were enrolled in our training program or took the test to find out if they wished to purchase licenses for its use. The test results had no important consequences for them individually.

REFERENCES

Bagby, R. M., Parker, J. D. & Taylor, G. J. (1994). The twenty item Toronto Alexithymia Scale-i. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*, 23–32.

Barger, V. A. (2012). *Individual differences in need for approval: Measurement and marketing implications*. Ann Arbor, MI: ProQuest Information & Learning.

Bäccman, C., Sjöberg, L. & Almqvist, K. (In press). A comparison of applicants' and incumbents' mean scores on health constructs and personality constructs. A follow-up study of military recruits in a selection setting. *International Journal of Selection and Assessment*.

Bergman, D., Sjöberg, L., Lornudd, C. & von Thile Schwartz, U. (2014). Leader personality and 360-degree assessments of leader behaviour. *Scandinavian Journal of Psychology*, *55*, 389–397.

Christie, R. & Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R. & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment*, *16*, 155–169.

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting and Clinical Psychology*, *24*, 349–354.

Donovan, J. J., Dwight, S. A. & Schneider, D. (2014). The impact of applicant faking on selection measures, hiring decisions, and employee performance. *Journal of Business and Psychology*, *29*, 479–493.

Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S. & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology*, *97*, 866–880.

Fine, S. & Pirak, M. (2015). Faking fast and slow: Within-person response time latencies for measuring faking in personnel testing. *Journal of Business and Psychology*.

Goffin, R. D. & Christiansen, N. D. (2003). Correcting personality tests for faking: A review of popular personality tests and an initial survey of researchers. *International Journal of Selection and Assessment*, *11*, 340–344.

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

Holden, R. R. & Hibbs, N. (1995). Incremental validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality*, *29*, 362–372.

Jones, A. & Crandall, R. (1986). Validation of a short index of self-actualization. *Personality and Social Psychology Bulletin*, *12*, 63–73.

Khorramdel, L. (2014). The influence of different rating scales on faking in high stakes assessment. *Psychological Test and Assessment Modelling*, *56*, 154–167.

Kovačić, M. P., Galić, Z. & Andreis, L. (2014). Warning against faking on personality questionnaire: Are warned participants more honest? *Contemporary Psychology*, *17*, 35–52.

Levashina, J., Weekley, J. A., Roulin, N. & Hauck, E. (2014). Using blatant extreme responding for detecting faking in high stakes selection: Construct validity, relationship with general mental ability, and subgroup differences. *International Journal of Selection and Assessment*, *22*, 371–383.

Matthews, G. & Oddy, K. (1997). Ipsative and normative scales in adjectival measurement of personality: Problems of bias and discrepancy. *International Journal of Selection and Assessment*, *5*, 169–182.

Mehrabian, A. & Epstein, N. (1970). A measure of emotional empathy. *Journal of Personality*, *40*, 525–543.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, *77*, 531–552.

Miller, B. K. & Ruggs, E. N. (2014). Measurement invariance tests of the faking sub-scale of the balanced inventory of desirable responding. *Personality and Individual Differences*, *63*, 36–40.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K. & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683–729.

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.

Ortner, T. M. & Schmitt, M. (2014). Advances and continuing challenges in objective personality testing. *European Journal of Psychological Assessment*, *30*, 163–168.

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

Paulhus, D. L. (2012). Overclaiming on personality questionnares. In M. Ziegler, C. Maccann & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 151–164). New York: Oxford University Press.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research. Explanation and prediction*. New York: Holt, Rinehart and Winston.

Reeder, M. C. & Ryan, A. M. (2012). Methods for correcting for faking. In M. Ziegler, C. Maccann & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 131–150). New York: Oxford University Press.

Robson, S. M., Jones, A. & Abraham, J. (2008). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance*, *21*, 89–106.

Roulin, N., Bangerter, A. & Levashina, J. (2014). Interviewers' perceptions of faking in employment interviews. *Journal of Managerial Psychology*, *29*, 141–163.

Schmitt, N. & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology*, *91*, 613–621.

Schutte, N. S., Malouff, J. M., Hall, L. E., Haggerty, D. J., Cooper, J. T., Golden, C. J., *et al*. (1998). Development and validation of a measure of emotional intelligence. *Personality and Individual Differences*, *25*, 167–177.

Sjöberg, L. (2010). *UPP-testet: Mångfald gynnas av korrektion för skönmålning*. [The UPP test: Diversity profits from correction for faking]. *Forskningsrapport 2010: 2*. Stockholm: Psykologisk Metod AB.

Sjöberg, L. (2014). *UPP-testet. Teknisk manual*. [The UPP test. Technical manual]. Stockholm: Psykologisk Metod AB.

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.

Sjöberg, L., Svensson, E. & Persson, L.-O. (1979). The measurement of mood. *Scandinavian Journal of Psychology*, *20*, 1–18.

Sjöberg, L. & Wolgers, G. (2012). *Personlighetstestning vid antagning till officersutbildningen* [Personality testing for the selection to Officer Cadet School]. ISSL Report F: 39. Karlstad: National Defence College.

Stanush, P. L. (1997). *Factors that influence the susceptibility of self-report inventories to distortion: A meta-analytic investigation*. Ann Arbor, MI: ProQuest Information & Learning.

Stark, S., Chernyshenko, O. S., Drasgow, F., Nye, C. D., White, L. A., Heffner, T., *et al*. (2014). From ABLE to TAPAS: A new generation of personality tests to support military selection and classification decisions. *Military Psychology*, *26*, 153–164.

Vecchione, M., Dentale, F., Alessandri, G. & Barbaranelli, C. (2014). Fakability of implicit and explicit measures of the big five: Research findings from organizational settings. *International Journal of Selection and Assessment*, *22*, 211–218.

Weinryb, R. M., Gustavsson, J. P., Åsberg, M. & Rössel, R. J. (1992). The concept of alexithymia: An empirical study using psychodynamic ratings and self-reports. *Acta Psychiatrica Scandinavica*, *85*, 153–162.

Wiggins, J. S. & Trapnell, P. D. (1997). Personality structure: The return of the Big Five. In R. Hogan, J. Johnson & S. Briggs (Eds.), *Handbook of personality psychology* (pp. 737–766). San Diego, CA: Academic Press.

Zickar, M. J. & Sliter, K. A. (2012). Searching for unicorns. Item response theory-based solutions to the faking problem. In M. Ziegler, C. Maccann & R. D. Roberts (Eds.), *New perspectives on faking in personality assessment* (pp. 113–130). New York: Oxford University Press.

Ziegler, M., Maccann, C. & Roberts, R. D. (2012). *New perspectives on faking in personality assessment*. New York: Oxford University Press.

Received 22 October 2014, accepted 13 April 2015

APPENDIX

The method used for correcting personality scales for faking is described in this Appendix.

The formulas will be most easily grasped if the raw values have first been standardized to mean = 0 and standard deviation = 1. For the case of only one independent variable, the regression coefficient equals the correlation between the dependent and independent variable; with two or more independent variables, the same principle applies but the formulas are more complicated. All this information is available in standard textbooks on statistics, see for example, Cohen, Cohen, West and Aiken (2003) for a comprehensive and advanced discussion. Explicit formulas are rarely provided for the case with two predictor variables, but they may be found e.g. in Pedhazur (1982).

The model for a test variable y (in z form, i e standardized to mean = 0 and standard deviation = 1) and two measures of faking, a and b (both in z form) is as follows:

$$y = \beta_a a + \beta_b b + e$$

where: β_a and β_b are the standardized regression weights; y, a and b are z-transformed forms of the dependent variable (personality scale) y; a and b are the two faking scales (Overt and Covert); e is an error term.

β_a and β_b are the standardized regression weights, estimated as follows:

$$\beta_a = (r_{ya} - r_{yb} \cdot r_{ab})/(1 - r_{ab}^2)$$

$$\beta_b = (r_{yb} - r_{ya} \cdot r_{ab})/(1 - r_{ab}^2)$$

where: r_{ya} the correlation between the test variable y and Overt faking a; r_{yb} is the correlation between test variable y and Covert faking b; and r_{ab} is the correlation between the Overt and Covert faking scales.

The expected value of the dependent variable y predicted from a and b is obtained by

$$y_{pred} = \beta_a a + \beta_b b$$

The residual is

$$res_y = y - y_{pred}$$

This residual is the test value corrected for faking, with simultaneous correction for Overt and Covert faking.

The procedure can of course be performed in a statistical program such as SPSS; it can also be programmed explicitly using the above formulas. This methodology utilizes faking scales maximally, under the assumption of linear regressions and equal regression weights under faking and non-faking.

© 2015 Psykologisk Metod AB. Scandinavian Journal of Psychology published by Scandinavian Psychological Associations and John Wiley & Sons Ltd.