Review of Psychology,
2009, Vol. 16, No. 2, 115-121

UDC 159.9

Three more attempts to prevent faking good in personality questionnaires

KLAUS D. KUBINGER

This paper describes the attempt to prevent faking good in personality questionnaires by several (new) means. Firstly, an analog response format was used instead of forced choice or another categorical response format; results based on the replication of an earlier experiment are presented. Secondly, the hypothesis that an "over-kill" number of items cause an examinee to finally give up faking good was considered. Thirdly, the hypothesis was tested that faking good can be prevented by use of a warning instruction stating that the computer is able to identify whether or not an examinee's answers fit a realistic personality profile. Both the latter hypotheses were investigated and rejected in another experiment. The same is true as concerns the replication of the experiment applying an analog response format. However, it is argued that the experiments described in this paper use volunteers as subjects; evidence is given that answering behavior changes considerably, depending on whether an examinee is a job applicant or merely a volunteer.

Key words: personality questionnaire, faking good, faking instruction, analog response format, experiment

There is a good deal of evidence showing that personality questionnaires are fakeable. Impressive evidence was first given by Viswesvaran and Ones (1999), but also since then, in particular by Birkeland, Manson, Kisamore, Brannick, and Smith (2006), Deller, Ones, Viswesvaran, and Dilchert (2006), and Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007a). In this paper, the phenomenon of faking good – especially in the context of job recruitment – is of prime interest, though faking bad also occurs within a clinical context (for a review, cf. Franke, 2002). Most studies on this topic use some kind of instruction to fake, which might be the reason why practitioners do not take these results into account. They argue that experimental behavior is different from behavior within a job recruiting procedure. Sometimes it is even argued that faking indicates a kind of social competence and is beneficial within recruiting (cf. for instance Marcus, 2003), or that faking is not harmful because everybody fakes and thus only the mean score is altered (cf. Kanning & Holling, 2001). Of course, there is no reason to trust so-called lying scales, given that these have proven to be fakeable as well (cf. Moorman & Podsakoff, 1992). Although it is easy to counter these claims with various arguments, their discussion is of little relevance to this paper and only two such arguments will be mentioned. Firstly, social competence in no way guarantees that the original personality traits, which the questionnaire aimed to measure, are still given satisfactorily, while loyalty to the company is undermined. Secondly, the differing behavior of different people is disregarded and therefore candidates who do not fake or who fake less are at a disadvantage.

The question is no longer whether examinees actually fake in answering a questionnaire, but rather what psychological means are at psychology's disposal to conceptualize personality questionnaires that are – for the most part – able to prevent an examinee from faking. In answering this question, the results of two experiments will be presented.

EXPERIMENT I:
The use of an analog scale response format

Instead of a traditional response format, such as the dichotomous (forced choice) response format in particular, an analog scale can be used in personality questionnaires. In this way, an examinee has the freedom to grade his/her answer to the given statement of an item on a continuum, bordered by two extreme verbalizations ("yes" vs. "no"; "right" vs. "wrong", and the like). If a computer is used for test administration, even then at least 150 (invisible) pixels on a line provide an almost infinite number of optional answer categories. Applying both analog scale response format and dichotomous response format, Karner (2002) established an

Klaus D. Kubinger, Division of Psychological Assessment and Applied Psychometrics, Center of Testing and Consulting, Faculty of Psychology, University of Vienna. Liebiggasse 5, A-1010 Vienna, Austria.
E-mail: klaus.kubinger@univie.ac.at (the address for correspondence).

important phenomenon: psychometric analyses of the same questionnaire (MBTI; Myers-Briggs Type Indicator – German edition, Bents & Blank, 1995) disclosed that when administered with the original dichotomous response format, the questionnaire's scales failed to fit the Rasch model. When, however, the test was administered with an analog scale response format and the responses subsequently were dichotomized, the Rasch model held. This meant that a large number of items were necessarily deleted (in order to establish at least *a-posteriori* Rasch model fit) in the former case, and almost no items in the latter case. Although the study did not refer specifically to faking, it did indicate that a dichotomous response format is likely to produce measures which are not unidimensional. The opposite is the case with an analog scale response format, that is, on the whole, this format actually measures unidimensionally. One can think of: the dichotomous response format probably encourages an examinee to use a specific answering strategy, so that the resulting score is not a fair indication of the characteristic which was intended to be measured. In other words, only the freedom to grade his/her responses seems to motivate an examinee to answer homogeneously.

A consequent hypothesis is that the dichotomous response format offers no challenge to the examinee if he/she tries to fake, but that the analog scale response format does make it difficult to strategically decide which particular grade of response sounds truthful but nevertheless tends towards what is socially desirable (personal advantage). A faking instruction experiment by Kubinger (2002) actually established a significant interaction effect for one of the four scales ("agreeableness") from a Big Five-like adjective list questionnaire. The questionnaire was administered to 151 psychology students – the faking instruction was "to imagine the challenge of undergoing a university admissions test for psychology," the neutral instruction was "to be aware that personality inventories are only useful if an examinee responds truthfully." The mean score within the randomized group to which the dichotomous response format had been administered was lower when examinees were additionally (randomly) confronted with the faking instruction than when they were confronted with the neutral instruction. The opposite was true for the group to which the analog scale response format had been administered: examinees with the faking instruction produced a higher mean score than when they were confronted with the neutral instruction. As a low score indicates a high degree of agreeableness, this result confirms the faking good phenomenon for the group with a dichotomous response format. However, given the analog scale response format, the resulting means of scores completely contradict this phenomenon. The interpretation of these results suggests that examinees have the tendency to evaluate themselves in a way that is in keeping with their character when rating themselves on the analog scale, whereas they deny having the respective attitude when confronted only with the dichotomous question. This indicates an effective way of preventing faking good. One must, however, bear in mind that only a single scale happened to show a significant effect that supports this expectation.

METHODS

We attempted to replicate that experiment, primarily because one of four scales do not necessarily mean a generalized effect, but might be a matter of type-I-error. The hypothesis is that rather the dichotomous response format than the analog scale response format brings an examinee to fake a personality questionnaire. A Big Five-like adjective list questionnaire, B5PO (*Big Five Plus One Personality Inventory;* Holocher-Ertl, Kubinger, & Menghin, 2003), which encompasses an additional sixth scale "empathy," was used. The experimental design and the instructions remained unaltered. The population was again made up of psychology students. Sixty six 3rd year students were all tested at one time during a psychological assessment lecture. The test was presented as having a didactical purpose. There was no way out to complete the paper-and-pencil questionnaire. They were all advised three times to read the instruction carefully. Asking students after the test administration – again during the psychological assessment lecture – whether they had actually tried to act according to the different instructions (but did not forget them while answering the questionnaire) proved that the experimental manipulation had prevailed; only two examinees did not raise one's hand.

RESULTS

Multivariate analysis of variance shows a significant main effect of instruction (*Hotelling's trace:* p = .015), however, no significant interaction occurs between the instruction and response format (*Hotelling's trace:* p = .732), and the significant main effect refers exclusively to the scale "agreeableness" (F = 7.841, p = .007). The difference in the means of the scores between the two groups with different instructions amounts to 1.02 by a standard deviation of 3.36, which is an estimated relative effect size of 0.30. The direction of mean difference confirms the faking good phenomenon.

DISCUSSION

This means that there is not only no general effect of the analog scale response format, but our experiment contradicts even Kubinger's (2002) results: At the moment, the analog scale response format does not seem capable of preventing faking good in personality questionnaires. Yet though the estimated relative effect size is rather small and applies only to a single scale, faking good instruction is proven once more to work.

SOME BY-THE-WAY EXPERIMENT: The use of a large number of items

The relevant research topic of preventing faking good in personality questionnaires by using an analog scale response format led to the question of whether fakeability is given in questionnaires for children as well. Seiwald (2002) investigated this question experimentally. Every child (10 to 14 years) in the sample was given both response formats, which were randomized in such a way that the first half of all 210 items were in the dichotomous scale format and the second half were in the analog scale format or vice versa – the halves were balanced. Once again, a faking instruction was used; that is, a 2×2×2 design was applied. Apart from the unexpected fact that the children scored higher in terms of social desirability when using the analog scale response format but not in the case of the dichotomous response format (in 16 out of 17 scales), the analog scale also yielded higher average socially desired scores on items that were given in the first half of the questionnaire. Two possible explanations are: Either children forget the faking instruction while answering the questionnaire (this would mean a serious objection to any experiment that uses a faking instruction), or children only get tired of giving socially desired answers if the number of items is very large.

As a consequence, a new hypothesis is now raised: Faking good may be prevented by a very large number of items, insofar as that the scores of the items at the end of a questionnaire correctly reflect the traits aimed to be measured.

Hence, a second experiment was designed in order to test this hypothesis; it will be described further on in this article. There is, however, a third means of preventing faking good in personality questionnaires that should be reflected upon first.

EXPERIMENT II: The use of a warning instruction

Although Menghin and Kubinger (1996) have long established that the use of a computer instead of test administration in person does not have any effect on preventing faking good, the computer nevertheless offers the possibility of combating the faking good phenomenon on personality questionnaires. That is, apart from the known effect of some warning instructions (cf. Dwight & Donovan, 2003; McFarland, 2003; Vasilopoulos, Cucina, & McElreath, 2005), it is nowadays actually feasible to use the computer to calculate whether any given answer of an examinee is likely or not: Psychometrics has developed certain person fit indices originally intended to identify examinees who master a Rasch model fitting achievement test by improper means. This works by calculating the probability of an examinee's actually solving a certain item, given his/her performance on previous items and given the item difficulty. In the meantime, this approach has been tested for the purpose of identifying examinees who fake a personality questionnaire (cf. Ponocny & Klauer, 2002). Whether or not this approach offers the ultimate solution to the problem of faking good, examinees may be leery of faking if they are made aware of this psychometric possibility of discovering that they have faked. Hypothetically speaking, most people believe in all kinds of computerized possibilities.

In our case, we used the following warning instruction: "You will be given a large number of items to answer, because this enables the computer to check whether your answers actually fit a realistic personality profile. This basically means that faking does not pay off. If faking is suspected, then you will be asked to work through the questionnaire again." The experiment indicated in the previous chapter was designed in accordance with this warning instruction.

METHOD

Firstly, a personality questionnaire was compiled that consisted of a large number of items, altogether 466 items, 67 of which were repeated. Although all questionnaires and scales were from German editions, they resembled state of the art international personality questionnaires: FAF ("aggression"; Hampel & Selg, 1975), FSKN ("self-esteem"; Deusinger, 1986), FKK ("locus of control of reinforcement"; Krampen, 1991), FBS ("tendency to commit suicide"; Stork, 1972), PD-S ("paranoia" and "depression"; von Zerssen & Koeller, 1976), TPF (a 9-scales personality inventory based on the construct of psychical health; Becker, 1989). That is to say, all the questionnaires as a pool seem to be representative for all international questionnaires.

A six category response format (graded from "completely true" to "completely false") was used for all items except the very last 40, which were presented in the dichotomous response format and were all repeated items. In this experiment, no faking instruction was given; the single experimental factor was that an additional warning instruction was either given or not.

The subjects were once again psychology students fulfilling a test experience requirement for their course. The experiment was carried out in accordance with a type-I-risk of 5%, a type-II-risk of 20% and a relative effect of d = .80. For the (one-sided) t-test, CADEMO (cf. Rasch & Kubinger, 2006) calculated a minimum sample size of n = 20 in each of the two independent experimental groups (with and without the given warning instruction). In effect, a total of 39 students provided useful data. All answers were dichotomized.

The main hypothesis was: "The warning instruction leads to a less socially desired mean of scores than is the case if the warning instruction is not given". This applies to the scale with the highest face-validity in terms of social desirability; it was decided that this scale should be "spontaneous aggression". Of course, similar analyses with respect to the other scales were also carried out, but in order to fix the type-I-error, this single scale became the critical scale: This means that if the *t*-test results in significance, the warning instruction is proven to pay off; whether or not additional scales confirm this result is then of less importance. Conversely, if the *t*-test does not result in significance with respect to the scale "spontaneous aggression," we need not look for other significances, because this would mean risking a high type-I-error. The interpretation of significant results might consequently be based on a statistical artifact. The hypothesis is restricted to the first half of the items due to the possibility that the mentioned effects of forgetting and tiredness do actually apply.

The second hypothesis, concerning the effects of forgetting and tiredness, is: "Questions administered twice, once at the beginning of the questionnaire and once again at the end, are answered in a less socially desired manner the second time around." To test this hypothesis, a potentially changing response behavior should be measured item-wise; the hypothesis would be supported if changes from "completely true" to "completely untrue" and vice-versa are not balanced and if changes towards answering according to social desirability occur more frequently. The McNemar test may be applied to test significance by summing up changes counted from all items whose contents are judged as being most indicative of certain social desires.

If the changes are considered separately for the groups with and without the warning instruction, then this computation also refers – although only descriptively – to the main hypothesis. If the number of changes indicating a tendency away from social desirability is larger in the group with the warning instruction than in the group without the warning instruction, then this instruction must have fulfilled its purpose, though only at the beginning of the questionnaire.

Finally, due to the fact that the repeated items are administered with either a six-category response format or the dichotomous format at the very end of the questionnaire, the comparison of changes towards social desirability within both of these response formats would once again indicate whether the dichotomous format differs from other formats with respect to faking good.

RESULTS

The result of the *t*-test, with regard to "spontaneous aggression" was not significant (t = 1.58; df = 35; p = .123). The respective items are located on average at position 4.70 to 5.02 – this tendency lies in accordance with the hypothesis. For further information, see Table 1, in which the other scales positioned in the first half of the questionnaire are presented according to the comparison of their mean scores.

Table 2, below, shows the results according to the second hypothesis as follows: For every repeated item, the number of changes from "completely true" to "completely untrue" and vice-versa are given. This applies to both the overall sample as well as the two experimental groups, with and

Table 1
Mean scores of scales from personality questionnaires in two experimental groups - one group with the warning instruction that faking good does not pay off and the other group without a warning instruction

Scale	With warning instruction	Without warning instruction	Significance*
Spontaneous aggression	89.4	95.5	.123
Reactive aggression	57.1	59.3	.355
Excitability	52.4	51.0	.563
Self aggression/ depression	46.0	46.5	.869
Inhibition of aggression	35.9	34.8	.569
Control of behavior	58.9	58.9	.973
Psychological health	58.0	51.4	.147
Achievement of sense	36.9	37.2	.922
Self abandonment vs. self centered	33.6	33.3	.886
Freedom of complaints vs. anxiety	32.0	30.5	.605
Expansivity	37.2	33.9	.245
Autonomy	49.6	45.9	.234
Feeling of self-worth	33.6	31.3	.298
Ability to show affection	28.1	24.8	.095
Risk of suicide	115.1	138.2	.293

Note. *Significance according to the *t*-test.

Table 2
Counted changes from “completely true” to “completely untrue” and vice-versa for a small sample of altogether 67 twice administered items

Position of the item	Item	Overall sample		With warning instruction		Without warning instruction	
Social desirability		Changed from “true” to “untrue”	Changed from “untrue” to “true”	Changed from “true” to “untrue”	Changed from “untrue” to “true”	Changed from “true” to “untrue”	Changed from “untrue” to “true”
6 /untrue	I quite honestly sometimes enjoy teasing other people.	**4**	**0**	**1**	**0**	**3**	**0**
12	(I am on my guard when I meet people who are friendlier than I expected.)	3	1	1	1	2	0
18 /untrue	I sometimes picture how awful people who have done me wrong must feel.	**8**	**2**	**4**	**0**	**4**	**2**
32 /true	I can not think of any valid reason for someone having to hit another person.	6	3	2	2	4	1
37 /untrue	I openly admit to having tortured animals before.	**2**	**0**	0	0	**2**	**0**
…	…	…	…	…	…	…	…
340 /untrue	I find it difficult to represent a group when faced with a conflictive opinion.	**7**	**3**	**4**	**2**	**3**	**1**
380 /true	I believe that I can justify my behavior in most cases.	6	6	4	4	2	2
382 /untrue	I have difficulties doing what is right.	4	7	2	6	2	1
396 /true	I am satisfied with myself.	**3**	**4**	3	1	**0**	**3**
Sum		193	145				

Note. Bold numbers indicate that the relationship between changes from first to second response is in accordance with a tendency to answer in a socially desirable manner. The sums are given for both the overall sample, as well as for the samples with and without the warning instruction. Apart from the position of the items at the first time of administration, the suggested answers which would correspond with social desirability are given.

without the warning instruction. In cases where the content of an item cannot be judged in terms of any social desirability, these items are set in brackets. If the relationship between changes from the first to the second response for the same item is in accordance with a tendency towards social desirability, this is also indicated. Finally, a row of summed counts is given for items whose contents are very likely to appeal to certain social desires. The sums of these counts were used to apply the McNemar test.

The McNemar test resulted in significance ($z = 2.6109$; $p = .0045$). On the one hand, the corresponding effect of 57.1 percent, instead of 50 percent, in favor of our hypothesis is neither impressive nor convincing. On the other hand, and more importantly, calculating the McNemar test for only those items that continued using the six-category response format for both administrations leads to a non-significant result. Over and above this, the changes go in the wrong direction ($z = -0.6030$). Therefore, the mentioned significant effect only refers to those items whose response formats were modified from the six-category response format to the dichotomous response format: $z = 4.3989$, the effect being a percentage of 67.3, instead of 50 percent, in favor of our hypothesis. This means that when the second time the dichotomous response format is used instead of the format with six categories, examinees give significantly more socially desirable answers. Finally, in viewing the results in Table 2 from a qualitative perspective with respect to both experimental groups, no unequivocal tendency is disclosed.

DISCUSSION

Firstly, the main hypothesis is not confirmed. Not a single scale leads to a significant *t*-test. The non-significance with respect to the target scale (“spontaneous aggression”) implies that the warning instruction was not effective. This interpretation is supported by the evaluation of changes towards social desirability on the repeated items; there was evidently no difference between the two experimental groups. We are not of the opinion that this result had anything to do with the actual formulation of the warning instruction; nevertheless, we shall discuss the possibility of using warning instructions in order to prevent faking good in the next chapter.

Secondly, the hypothesis that examinees get tired or give up faking good when administered with an “over-kill” number of items is also rejected. We do not believe that such an effect could be established if the number of items was increased further, because 466 is already an almost unreasonable number of items. In other words, this hypothesis should be completely disregarded in future.

Thirdly, the dichotomous response format again showed itself to be disadvantageous with regard to the faking good phenomenon. It is obviously easiest to fake towards social desirability if the examinee is forced to choose only between two extreme answers.

The main problem of research work on the phenomenon of faking personality questionnaires is most certainly the kind of population from which the subjects are sampled. Like this paper, many other studies and experiments use volunteers. It is, however, highly likely that volunteers behave quite differently than people who are tested within a job recruiting process; see for instance Birkeland et al. (2006) or, for a psychometrically stronger approach, Karner (2002): Using an ex-post-facto experiment comparing volunteers and selection candidates, he disclosed that while the considered scales appear to stand Rasch model checks for voluntary subjects, i.e. they are all likely to measure unidimensionally in this population, this is not at all the case for examinees in a selection scenario. Regardless of whether the response format is a dichotomous or an analog scale or whether the administration is carried out in person or by a computer, the items of the questionnaire fulfill psychometric presuppositions of a proper scale in the case of volunteers. In other words, personality questionnaires do their job well in the case of voluntary examinees; they however fail to measure fairly when consequences are involved for the examinee (cf. recently Morgeson et al., 2007b).

Hence, experiments on potential ways of preventing faking good on personality questionnaires require a population of selection candidates rather than volunteers. Forthcoming research should seriously take this into account. On the other hand, the irrelevance of results based on volunteers does give hope that a warning instruction may still work if selection candidates were to be tested.

CONCLUSION

So far, all attempts to find a means of preventing or at least considerably reducing the faking good phenomenon in personality questionnaires have failed. There is some hope that an analog scale response format sometimes works to prevent this phenomenon. Yet there is hardly a hope that an "over-kill" number of items is effective. There is also only a small ray of hope that a warning instruction (that it is better not to fake) may work. Nevertheless, the demand is made on psychology to test new, different means that may possibly help to overcome the problem (for instance, recently Khorramdel & Kubinger, 2006, investigated in particular the respective effect of speededness).

Of course, our results are restricted to volunteers (and the use of faking instruction); and we gave evidence that answering behavior changes considerably, depending on whether an examinee is a job applicant or merely a volunteer. Hence further research is needed concerning the relevant population of selection candidates.

REFERENCES

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.

Becker, P. (1989). *Der Trierer Persönlichkeitsfragebogen (TPF)* [*Trier Personality Questionnaire*]. Göttingen (Germany): Hogrefe.

Bents, R., & Blank, R. (1995). *Myers-Briggs Typenindikator (MBTI)* [Myers- Briggs Type Indicator] (2. ed.). Göttingen (Germany): Beltz.

Deller, J., Ones, D.S., Viswesvaran, C., & Dilchert, S. (Eds.)(2006). Considering response distortion in personality measurement for industrial, work and organizational psychology research and practice. Special issue: *Psychology Science, 48,* 1.

Deusinger, I.M. (1986). *Die Frankfurter Selbstkonzeptskalen (FSKN)* [*Frankfurt Self-Image Scales*]. Göttingen (Germany): Hogrefe.

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

Franke, G.H. (2002). Faking bad in personality inventories: Consequences for the clinical practice. *Psychologische Beiträge, 44,* 50-61.

Hampel, R., & Selg, H. (1975). *Fragebogen zur Erfassung von Aggressivitätsfaktoren (FAF)* [*Aggression Factors Questionnaire*]. Göttingen (Germany): Hogrefe.

Holocher-Ertl, S., Kubinger, K.D., & Menghin, S. (2003). *Big Five Plus One Persönlichkeitsinventar (B5PO)* [*Big Five Plus One Personality Inventory*]. Test: Software and Manual. Mödling (Austria): Wiener Testsystem/ Schuhfried.

Kanning, U.P., & Holling, H. (2001). Struktur, Reliabilität und Validität des NEO-FFI in einer Personalauswahlsituation [Structure, reliability, and validity of NEO-FFI at personnal selection]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 22,* 239-247,

Karner, T, (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge, 44,* 42-49.

Khorramdel, L., & Kubinger, K.D. (2006). The effect of speededness on personality questionnaires – An experiment on applicants within a job recruiting procedure. *Psychology Science, 48,* 378-397.

Krampen, G. (1991). *Fragebogen zu Kompetenz- und Kontrollüberzeugungen (FKK)* [Questionnaire of Competence and Control Persuasion]. Göttingen (Germany): Hogrefe.

Kubinger, K.D. (2002). On faking personality inventories. *Psychologische Beiträge, 44,* 10-16.

McFarland, L. A. (2003). Warning against faking on a personality test: Effects on applicant reactions and personality test scores. *International Journal of Selection and Assessment, 11*, 265-276.

Marcus, B. (2003). Das Wunder sozialer Erwünschtheit in der Personalauswahl [The miracle of social desirability at personnal selection]. *Zeitschrift für Personalpsychologie, 2*, 129-132.

Menghin, S., & Kubinger, K.D. (1996). Zur Legende: „Testpersonen beantworten dem Computer persönliche und intime Fragen offener als einem Testleiter" - Ergebnisse eines Experiments [On the myth of whether examinees answer intimate questions more truthfully if faced with a computer rather than an administrator in person – results from an experiment]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 17,* 163-169.

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

Morgeson, F.P., Campion, M.A., Dipboye, R.L., Hollenbeck, J.R., Murphy, J.R., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60,* 1029-1049.

Moorman, R., & Podsakoff, P.M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65,* 131-149.

Ponocny, I., & Klauer, K.C. (2002). Towards identification of unscalable personality questionnaire respondents: The use of person fit indices. *Psychologische Beiträge, 44,* 94-107.

Rasch, D., & Kubinger, K.D. (2006). *Statistik für das Psychologiestudium - Mit softwareunterstützter optimierter Untersuchungsplanung: Stichprobenvorausberechnung und Sequentielles Hypothesenprüfen* [*Statistics for studying psychology - Including computer assisted optimal experimental design: Sample size determination and sequential testing of hypotheses*]. Heidelberg: Spectrum.

Seiwald, B.B. (2002). Replicability and generalizability of Kubinger`s results: Some more studies on faking personality inventories. *Psychologische Beiträge, 44,* 17-23.

Stork, J. (1972). *Fragebogentest zur Beurteilung der Suizidgefahr (FBS)* [*Suicide Questionnaire*]. Salzburg (Austria): Müller.

Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 90,* 306-322.

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

Zerssen, von D., & Koeller, D.M. (1976). *Paranoid-Depressivitäts-Skala (PD-S)* [*Paranoia-Depression Scale*]. Weinheim (Germany): Beltz.