Psychologische Beiträge, Band 44, 2002, S. 24-41

Comparing verbal and non-verbal personality scales: Investigating the reliability and validity, the influence of social desirability, and the effects of fake good instructions

MANFRED AMELANG[1], ANDREAS SCHÄFER & SAFIR YOUSFI[2]

Summary

The fakeability of verbal and non-verbal personality questionnaires was investigated. Participants completed scales from the Personality Research Form (*PRF*, Jackson, 1967), the matching scales from the Non-verbal Personality Questionnaire (*NPQ*, Paunonen, Jackson & Keinonen, 1990) and a social desirability scale in two sessions. At first contact, all participants responded to standard instructions. At second contact, the condition for half of the participants was an imagined job-application (fake good treatment), the other half received standard instructions again (control). The non-verbal scales correlated less highly with criterial peer ratings and social desirability than the verbal questionnaires, but the differences were rather small. Under the faking condition, however, the *NPQ* and *PRF* scores were affected almost to the same extent. The validity of both scales was more or less evenly impaired, but reached coefficients above .40 even under the fake good condition. This implies that within the context of research conditions personality questionnaires may retain their (limited) peer-rating based validity despite faking effects.

Key words: social desirability, verbal and non-verbal questionnaires, fakability, peer validity

Introduction

Although self-report questionnaires are the most commonly used instrument in personality measurement, our field continues to be concerned with the issue of response biases. More than 30 years ago, Jackson and Messick (1967) conducted a series of factor analyses of *MMPI* items and put forward a provocative conclusion: response differences on the *MMPI* might reflect little more than individual differences in the tendency to answer according to social desirability or acquiescence. This extreme view is no longer held, even by its initial proponents, largely due to the fact that self-report measures of personality have accumulated a proven track record of external validities, especially with respect to peer ratings. Nonetheless, the sole reliance on the verbal medium remains a cause for concern since most personality questionnaires use verbally formulated items and present response options that are, at

[1] Prof. Dr. Manfred Amelang, Department of Psychology, University of Heidelberg, Hauptstraße 47-51, D-69117 Heidelberg, Germany

[2] We are grateful to Oliver John and to Kimberly Feldt for their very helpful suggestions to a former draft of the present paper. In addition we thank Birgit Koopmann for helping us with the editing of this paper.

the very least, "verbally anchored" - such as "Yes-No," "True-False" or "Agree-Disagree" to varying degrees.

Such verbally based ratings of personality traits have been criticized most severely by Shweder (1982) who proposed the "semantic distortion hypothesis." According to Shweder, correlations among trait ratings may reflect the semantic similarity among the trait words, rather than the empirical co-occurrence of trait relevant behaviors. Initial analyses seemed to provide some support for the operation of semantic similarity biases. However, subsequent findings, especially the studies by Borkenau and Ostendorf (1987a, b), demonstrated the role of methodological artifacts as well as the difference between the structure and the validity of trait ratings - the substantial validities of trait ratings could not be explained by the biased process suggested by the semantic distortion hypothesis.

Similar explanatory limitations apply to the related conception of "implicit personality theories" which are assumed to reflect attributes of the perceiver or personality rater, rather than the one being rated. According to Mirels (1982), these implicit theories are "illusory in nature" rather than being based on an accurate representation of social reality. This perspective of the trait rating process cannot explain, however, the series of findings demonstrating substantial consensus correlations among independent raters of the same target individual (cf. Gosling, John, Craik & Robins, 1998, for a review).

Nonetheless, the sole reliance on verbal measures continues to make the field of personality assessment vulnerable to attacks. Thus, alternative approaches are desirable and must be considered carefully. One possibility has been propagated over the years by Paunonen and Jackson. Parallel to Jackson's Personality Research Form (*PRF*; Jackson, 1984), they developed a non-verbal personality test in which the item content is presented in pictorial form, i.e. rather ingenious drawings of stick figures performing a variety of activities or illustrating various psychological states.

In their initial work, Paunonen and Jackson (1979) used these test items to study the nature of trait inferences from verbal and non-verbal material and found few differences. Subsequently, they assembled 136 of these pictorial items into the Non-verbal Personality Questionnaire (*NPQ*; Paunonen, Jackson & Keinonen, 1990) to measure 16 of the need constructs also measured on the *PRF*. Two examples are given in Figure 1; the scene on the left is from the "Thrill Seeking" scale (which measures the opposite of the earlier "Harm Avoidance" scale from the *PRF*), and the scene on the right is from the "Nurturance" scale.

The task of the participant is to imagine the situation shown in the drawing and "estimate the likelihood that he or she would engage in the type of behavior shown by the central actor, the one with the black hair", using a 7-point rating scale from 1 (= extremely unlikely) to 7 (= extremely likely). In this way, no verbal description of the behavioral meaning of the questionnaire's item needs to be given, even though the exact processes by which the test taker makes the probability judgments are not yet fully understood (and might still involve some degree of verbal mediation).

Studies of North American, Polish, Finnish, and German subjects (Paunonen, Jackson, Trzebinski & Foersterling, 1992; Spinath & Angleitner, 1995) have shown that the non-verbal scales have reasonable internal consistency reliabilities (averaging about .70). They converge to about .50 with their corresponding *PRF* verbal scales. Moreover, these scales have a factorial structure that resembles the Five-Factor Model (Costa & McCrae, 1988; 1990) and it remains essentially invariant across these four Western cultures.

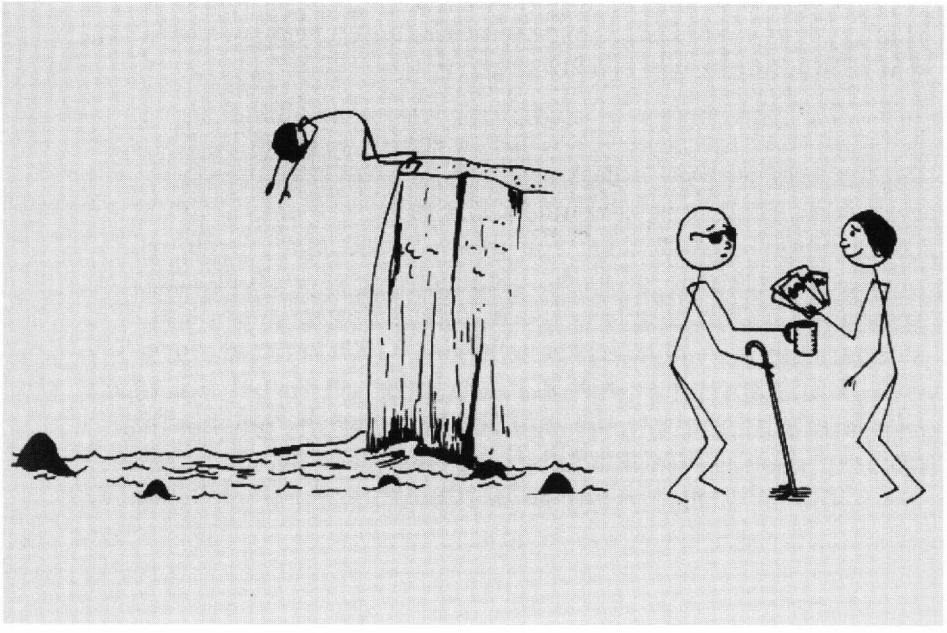

Figure 1

For a subset of *NPQ*-items constituting a "Five-Factor Non-verbal Personality Questionnaire" (*FF-NPQ*) Paunonen, Ashton and Jackson (2001) reported a self-peer correlation averaged across the five corresponding factor scales of $r = .41$. In a study with German subjects and using peer ratings as external criteria (one peer for each subject) Spinath and Angleitner (1995) reported for all *NPQ* scales a mean validity of .52.

The major goals of the present research are the following: We compare the new non-verbal questionnaire to the verbally based *PRF*, focusing on the reliability and validity compared to peer ratings, and then examine the effect of a fake good instruction in an experimental design. We test whether the scores on the non-verbal questionnaire are less susceptible to faking tendencies than the *PRF* and whether the non-verbal scales retain greater validity in the fake condition than the verbal *PRF* scales. More generally, the scientific utility of the non-verbal questionnaire will be tested and several general issues will be addressed regarding social desirability, both as manipulated by a fake good instruction and as measured as an individual difference variable.

One potential difference between the *NPQ* and the *PRF* is that the exact behavioral meaning of the *NPQ*-items is implicit. Like other "projective" tests that use pictorial stimuli, the respondent has to decide which behaviors are expressed by the drawings. There remains some uncertainty regarding the meaning of the items and therefore judgements about their desirability should be less clear and individuals should be less inclined to fake or present themselves in a particular way. In other words, just as projective tests are difficult to fake

because the individual does not know what is being measured, the non-verbal test may be more difficult to fake because item meanings are less clear.

These considerations suggest that:

- First, scales scores of a non-verbal questionnaire should be less highly correlated with social desirability response tendency than a verbal questionnaire.
- Second, the non-verbal questionnaire should be less susceptible to mean level changes in a fake good situation.
- Third, the reduction in validities likely to occur in a fake situation should be less pronounced in a non-verbal questionnaire than in a verbal questionnaire.

Method

Participants

A total of 190 participants (64% women), the average age being 37.1 years (*SD* = 12.8), provided self reports. This sample was recruited in four cities in Germany; participants were employed in a variety of occupations in sales, administrative, technical, and medical fields (no students).

Each participant also nominated two peers whom he/she had known for at least one year and who were well enough acquainted with him/her to provide a detailed personality description. Indeed, the peers (mean age 40 years, *SD* = 14) had known the participants for an average of 18 years (*SD* = 12). The peers returned their rating material directly to the experimenter, typically by mail.

Experimental design

The design of the study is outlined in Table 1.

In Session 1, the 190 participants completed the non-verbal and verbal personality questionnaires as well as a social desirability scale, under standard self-description instructions. For Session 2, subjects were randomly assigned to two groups, a control and an experimental group. One week after the first session, the control group again completed the same instruments under the same standard instructions. The experimental group, however, received a "fake good" instruction. These participants were asked to imagine themselves applying for an important job:

"When you respond to the questionnaire, please imagine yourself in the following situation: You have applied for a job that is very important to you and now you have to complete this questionnaire for your application. Your responses to these questions could make the difference whether or not you get the job. Please keep this in mind as you work through the questionnaire".

Table 1: Design of the study: self-ratings, peer ratings, and faking manipulation in Session 2

Session 1	To assess reliability and validity in a large sample:	
	Self-ratings (N=190)	Ratings by 2 peers (for all N=190)
Session 2	Random assignment of subjects to:	
	Control group (N=95)	Experimental group (N=95)
	Standard instruction	"Fake good" instruction

Measures obtained from each participant (both self and peer) at each session: *NPQ*: Nonverbal Personality Questionnaire (Paunonen et al., 1990), *PRF*: Personality Research Form (Jackson, 1984), *(G)SD*: German Social Desirability scale (Amelang & Bartussek, 1970)

Measures

Three instruments were administered to each participant (both self and peer) at each session: *NPQ, PRF*, and a scale measuring social desirability.

In the *NPQ* (Paunonen et al., 1990), each scale consists of 8 pictorial items to be rated on a 7-point scale. Instead of the original *PRF* (Jackson, 1984) we used a carefully translated and validated German version (Stumpf, Angleitner, Wieck, Jackson & Beloch-Till, 1985). Each of the *PRF* scales consists of 16 true-false items.

Both the *NPQ* and the *PRF* aim to measure constructs derived from Murray's (1938) set of needs; the *NPQ* includes scales for 16 of the 20 need constructs measured by the *PRF*. For the present study, we selected 9 scales to ensure that all of the Big Five domains of personality description would be represented using the factor loading information provided by Paunonen et al. (1992, 1996).

In addition to these 9 personality constructs, we also administered a German language social desirability scale (Amelang & Bartussek, 1970). This carefully constructed 32-item scale was derived from a large pool of items, including all of the Marlowe-Crowne Need for Approval items, the *MMPI* and Eysenck Lie Scale items, as well as some newly written items. Items were first selected rationally to reflect either unlikely virtues (desirable but judged infrequent) or common failings (undesirable but judged frequent); the most internally consistent sub-set of these items were retained for the final scale. The resulting scale is internally consistent, has high retest reliability and shows substantial increases under fake good in-

structions. Scores are not related to extraversion and only modestly negatively to neuroticism. Factor analyses showed stable results, with two correlated factors (about .50) which, like those discussed by Paulhus (1986), may be interpreted as self-presentation (claiming unlikely virtues, all keyed true) and denial (rejecting common failings, all keyed false). In short, both item content and scale characteristics are similar to the commonly used social desirability scales on which this scale is based (for a review, see Paulhus, Fridhandler & Hayes, 1997). Item examples are "As far as I can think back, I have never had quarrels with family members" (keyed true) and "Sometimes I arrive late for work or an appointment" (keyed false).

Results

Reliability: internal consistency, retest stability, and inter-rater agreement

We first examined the reliability of the two tests.

Table 2: Comparison of the verbal and non-verbal tests on three reliability indices (averaged across the 9 Scales)

	PRF	NPQ
Internal consistency (alpha)		
Self-reports--Session 1	.69	.67
Self-reports--Session 2	.70	.74
Peer-reports	.80	.73
Retest stability of self-reports		
Session 1-2	.74	.73
(Control group)		
Peer Agreement		
Intra-class correlation	.61	.60
Correlation between self-reports under standard instruction and fake good instruction Session 1-2 (Experimental group)	.62	.59

Table 2 shows the result for the *PRF* in the first column and for the *NPQ* in the second column, and the coefficients are averaged across the nine scales on each test. Reliability is calculated using the following methods:

First, the internal consistency (Cronbach's alpha) coefficients were computed across the items in each scale. For the self-reports in Session 1 and 2, the average scale had an alpha of approximately .70 for both tests. For the peer-reports, the alphas were a little higher for the *PRF* scales than for the *NPQ* scales.

Second, for the participants in the *control group*, we could compute retest stability coefficients. Again, there were no differences between the two tests: both had average scale stabilities that exceeded .70. The correlations for the participants in the *experimental group*, i.e. comparing the self-reports in Session 1 under standard instructions with those in Session 2 under fake good instructions nearly reached the level of the retest stability coefficient (*NPQ:* .59, *PRF:* .62), once again showing small differences between the two tests. In fact, a significant decrease in the fake good condition compared with the control condition could only be observed for four scales of the *NPQ* (*Understanding* $p < .01$, *Exhibition* $p < .05$, *Thrill Seeking* $p < .05$, *Nurturance* $p < .05$) and three scales of the *PRF* (*Understanding* $p < .01$, *Exhibition* $p < .05$, *Social Recognition* $p < .05$). The correlation of the Social Desirability Scale between Session 1 and 2 was also significantly lower under fake good conditions ($r = .57$ *vs.* $r = .78$, $p < .01$).

Third, for the two peers who rated each participant we computed inter-rater agreement using the intra-class correlation from Shrout and Fleiss (1979; formulas1, 2). Again, these coefficients did not differ between the two tests, averaging .60 across the nine scales.

To conclude, the two tests showed remarkably similar reliabilities.

Verbal-non-verbal convergence[3]

How well did the two tests converge? For each of the 9 constructs, we computed the correlation between the *PRF* scale and the *NPQ* scale. These convergence correlations, averaged across the nine constructs, are shown in Table 3.

Table 3: Convergence between the verbal and non-verbal tests (averaged across the 9 scales)

	PRF – NPQ correlation
Self-reports--Session 1	.45
Self-Reports--Session 2	.47
Peer-reports	.56

[3] We use that term for the correlation between the verbal and non-verbal tests instead of the term "concurrent validity" because validity would make it necessary to discriminate between a predictor and a criterion. The verbal and non-verbal tests do not allow an analogous discrimination, i. e. verbal and non-verbal tests are equal with respect to that status.

The correlations are significant and substantial in size, for the self-reports in Sessions 1 and 2, as well as for the peer-reports, averaging about .50. However, a correlation of .50 indicates only a moderate degree of convergence. More simply, the constructs measured by the verbal and non-verbal tests are similar but they are *not* the same; the non-verbal test cannot replace the verbal test, and *vice versa*. In other words, we cannot claim that these two tests are *parallel* tests.

Are these findings unique to our sample, and to the particular set of scales we selected for this study? Paunonen et al. (1992) report mean convergence correlations for the complete set of 16 scales in four cultures; .53 in Canada, .50 in Finland, .47 in Germany, and .40 in Poland. Our results are, therefore, quite typical.

Self-peer validity

Table 4 summarizes the validity findings using the means of the two peer-ratings as the criterion.

Table 4: Validity comparison: predicting peer-ratings on PRF and NPQ from self-reports on the two tests in session 1 (means across the 9 scales and ranges)

		Mean peer-ratings	
Self-reports		PRF	NPQ
PRF			
	Mean	***.56***	.38
	(Range)	(.43-.71)	(.12-.58)
NPQ			
	Mean	.36	***.52***
	(Range)	(.22-.53)	(.43-.59)

Note. (*N*=190)

The self-reports on *PRF* and *NPQ* are given as rows and the peer-reports are given as columns, and the self-peer validity correlations within each test are set in bold italics. Thus, the boldface entry in the upper left corner of the table indicates that *PRF* self-ratings correlated .56 with *PRF* peer-ratings. This is the average value across the nine scales, the values in parentheses indicate the range, the lowest correlation was .43 and the highest .71. Our mean of .56 is quite similar to the mean peer validity of .52 for the full *PRF* in Paunonen and Jackson's (1985) dormitory room-mate rating study in Canada.

The self-peer validity for the *NPQ* scales was quite similar, averaging .52, with a range from .43 to .59.[4]

How well can we predict *NPQ* peer-ratings from *PRF* self-ratings, or vice versa? The cross instrument correlations are shown on the off diagonal in Table 4 because they assess validity across two data sources (self vs. peer) as well as across the two tests (*PRF* vs. *NPQ*). Thus, these correlations have to be lower than the within-test correlations shown on the diagonaly. Nevertheless, both diagonal correlations are significant and exceeded .35. Again, they were quite similar (.36 and .38), revealing no substantial differences between the two tests. In summary, both tests showed similar levels of peer validity, exceeding .50 for the within-instrument correlations and .35 for the cross instrument correlations.

Correlations with social desirability scale scores

Our first analysis focused on individual differences; we correlated participants' scores on the German language social desirability scale with their scores on the *PRF* and *NPQ* scales.

The results are summarized in Table 5.

On average, the *PRF* self-reported scales correlated .27 with social desirability scores and the *NPQ* scales correlated .21, again quite similar values.

How should we interpret these small but significant correlations? It is possible that they do not indicate biased self perceptions, but accurate perceptions. For example, results in the second row of Table 5 show that the participants' social desirability scores correlated .27 with their self-reports of *PRF* achievement, but also .25 with peer-reports of *PRF* achievement. That is, individuals scoring high on social desirability not only claimed greater striving for achievement in their self-reports but their peers agreed (within the limits of these small correlations)! Considering this problem, John and Robins (1994) have recently argued for a more refined measure of self-report bias that partials out all valid variance (as measured by the peer criterion). Using this peer partialled measure, the coefficients in the last row of Table 5 show that the social desirability bias component in self-reports was not large and again of similar size in the two instruments: the average self bias correlations were .21 for the *PRF* scales and .17 for the *NPQ* scales.

---

[4] One might argue that validity coefficients of that magnitude are basically insufficient, because a correlation of about .50 between a predictor and a criterion variable means only about 25 % explained variance, but they reflect the empirical reality (see for instance Amelang & Zielinski, 1997, pp. 255-256).

Table 5: Correlations with Social Desirability Scale values in Session 1 for peer-ratings, for self-ratings, and for self-ratings with the valid peer variance partialled out (i.e., remaining bias variance)

	PRF			NPQ		
		Self			Self	
Scale	Peer	Corr.	Peer part.	Peer	Corr.	Peer part.
Endurance	.32	*.44*	*.35*	.13	.14	.09
Achievement	.25	*.27*	.16	.04	-.03	-.05
Understanding	.04	.10	.16	.08	.06	.02
Aggression (R)	.26	*.47*	*.40*	.24	*.34*	*.27*
Succorance (R)	.02	.19	.20	.11	*.35*	*.33*
Nurturance	.14	*.26*	*.22*	.12	.15	.10
Soc. Recogn. (R)	.05	.20	.20	.14	*.26*	*.22*
Thrill Seeking (R)	.26	*.22*	.05	.12	*.26*	*.24*
Exhibition (R)	.26	*.30*	.17	.27	*.33*	*.23*
Average	.18	.27	.21	.14	.21	.17

Note. Scales have been ordered according to independent judgments of how much scale scores should change if the individuals were trying to "fake good". (R) indicates that this scale was keyed in the socially undesirable direction but is shown here reverse keyed (e.g. Lack of aggression). Bias correlations (self with peer partialled out) larger than .20 are set in italics. All correlations of .15 and greater are significant at $p < .05$.

Faking effects on mean levels: experimental manipulation

The interpretative difficulties inherent in the individual differences approach to social desirability led us to include an experimental manipulation in Session 2 of this study. As outlined above, the participants in the fake good condition had completed the same instruments under standard instruction in Session 1, permitting us to examine changes in their scales scores.

The first question was whether the manipulation worked. Because the social desirability scale was administered both in Session 1 under the standard self-report instructions and in Session 2, we could conduct a direct manipulation check. Indeed, the 2x2 ANOVA showed the predicted interaction between Session (1 vs. 2) and Condition (control vs. fake good instruction), $F(1.188) = 34.5$, $p < .001$. As expected, the participants in the fake good condition showed a significant increase over their initial scores (obtained under standard instructions), $t(94) = 9.2$, $p < .05$, and they also scored significantly higher than the participants in the control condition (who completed the social desirability scale again under standard instructions), $t(188) = 2.1$, $p < .05$. In fact, the change from pre-test to fake good condition was substantial, amounting to almost a full standard deviation unit ($d_E = .93$). However, a slight retest effect was measured in the control group ($d_C = .22$, $t(94) = 3.36$, $p < .01$).

Significant faking effects (= interaction: session x experimental condition) could also be observed for the *NPQ* and *PRF* via global repeated measures MANOVAs (see last row of Table 6). Distinct repeated measures MANOVAs for the scales with high and low social desirability revealed only a significant interaction term for the scales with high social desirability (row 6 in Table 6) but not for the group of scales with low implications for social desirability (the next to last row in Table 6). In detail the following scales revealed significant univariate faking effects for both, the *PRF* and the *NPQ*: *Endurance, Achievement* and *Understanding* (see Table 6). These are the scales with the highest social desirability "loadings" (see the rank order in Table 6). Faking effects for *Succorance* were only significant on the *NPQ,* whereas for *Social Recognition* only the *PRF* scale revealed significant effects. Small but statistically significant retesting effects in the control group were observed on the *PRF* for *Aggression* and *Thrill Seeking* (see the column labelled "Control" in Table 6). On the *NPQ* there were also small retesting effects in the control group for the *Aggression* and the *Nurturance* scale.

To show the effects of faking on the *PRF* and *NPQ* scale scores, we computed the increase in scale scores. To make these change scores comparable across the two tests, we expressed them in standard score metric. These values are shown in Table 6 separately for each construct and, at the bottom of the table, averaged across the nine scales.

The overall means indicate that the *PRF* scale scores increased on average by .37 of a standard deviation and the *NPQ* scales by .30 of a standard deviation. These findings provide only weak support for the hypothesis that the non-verbal test is less strongly influenced by socially desirable response tendencies than the verbal test.

If we compare the individual scales across the two tests, the *PRF* showed a stronger faking effect for 5 scales and the *NPQ* a stronger effect for 4 scales. Again, there were no systematic differences between the two tests.

Validity changes due to faking

Finally, we examined how the fake good manipulation affected the validity of the self-reports in Session 2. Here we can compare the validity coefficients obtained for the participants in the control condition (those who completed the instruments under standard instructions) with the validity coefficients obtained in the experimental condition (fake good condition).

Table 6: Increases in standardized scale scores (d) in the control group and due to faking in the experimental group and tests for the Session x Group interaction

	Experimental			Control			Interaction Session x Group		
Scale	Mean	NPQ	PRF	Mean	NPQ	PRF	Mean	NPQ	PRF
Endurance	.82***	.74***	.90***	.12	.16	.08	***	***	***
Achievement	.70***	.67***	.73***	.07	.09	.04	***	***	***
Understanding	.44***	.45***	.42***	-.01	.01	-.02	***	***	***
Aggression (R)	.41***	.45***	.37***	.23***	.27**	.18*			
Succorance (R)	.35***	.44***	.25**	.02	-.08	.11	**	***	
Mean (absolute)	.54***	.55***	.53***	.08	.09	.08	***	***	***
Nurturance	.17*	.09	.24**	.14*	.16*	.11			
Soc. Recognition	.13*	.06	.20*	-.08	-.11	-.04	*		*
Thrill Seeking	.08	.10	.05	.12	.09	.15*			
Exhibition	-.03	-.24***	.18*	-.05	-.09	-.01	*		
Mean (absolute)	.09**	.00	.17***	.03	.01	.05			
Total mean (absolute)	.33***	.30***	.37***	.06	.05	.07	***	***	***

Note. Scales have been ordered according to independent judgments of how much scale scores should change if the individuals were trying to "fake good." (R) indicates that this scale was keyed in the socially undesirable direction but is shown here reverse keyed (e.g. Lack of aggression). Repeated-measures MANOVA tests are given in the "Mean" columns and rows. The aggregated session x group interaction terms of different scales contribute to the MANOVA effects (in the last three columns) regardless of their direction.
* $p < .05$, ** $p < .01$,*** $p < .001$.

Table 7: Effects of fake good on validity: self-peer correlations validity under either standard or fake good insturctions in Session 2

Self-reports in Session 2		Standard self instuction		Fake good self instruction	
		Peer PRF	Peer NPQ	Peer PRF	Peer NPQ
PRF					
	Mean	***.52***	.35	*.44*	.31
	(Range)	(.32-.65)	(.06-.61)	(.27-.60)	(.05-51)
NPQ					
	Mean	.31	***.54***	.28	*.43*
	(Range)	(.12-.50)	(.37-.67)	(.08-.54)	(.21-.54)

Note. N=95 in each condition. Self-peer correlations within each instrument are set in italics.

The left side of Table 7 provides a close replication of our peer validity analyses from Session 1 (see Table 4): under standard self-report instructions, the validities were again above .50 for within instrument analyses and above .30 for across instrument analyses. Comparing the validities of the individual scales between Session 1 and Session 2 we found only one significant difference in the control group (r = .67 in Session 2 vs. r = .37 in Session 1, p < .001).

What occurred as the participants were faking it in Session 2? In comparison to the control condition, the validity under fake good condition was smaller in 6 scales of the *NPQ* and in 7 scales of the *PRF*. The right side of Table 7 shows the expected drop: .44 and .43 for the within instrument analyses and .28 and .31 for the across instrument analyses, and this drop was roughly the same for both tests. Note, however, that this drop was smaller than one might have expected. In fact, only the *NPQ* scales *Succorance* (p < .001) and *Social Recognition* (p < .05) had significantly lower validities when presented under fake good conditions in Session 2. That is, individual differences remained fairly valid even in the fake good condition.

Differences in individual stability and validity between verbal and non-verbal forms

As stated above, each participant had responded twice to each item. Moreover, for each individual item response there was a homologous score based on the judgments of the peers. Therefore, it was possible to calculate individual coefficients of retest stability (i.e. correlating individually the responses to the items in Session 1 with those from Session 2) and validity (i.e. correlating individually the self- with the peer-responses to the items) for each participant. Each of these individual coefficients was calculated across the total of the number of items in the questionnaire (i.e. 186 *PRF*- and 80 *NPQ*-items). The resulting coefficients were

correlated with one another to test whether the *interindividual* differences of the psychometric properties of the non-verbal and verbal forms of the tests (and the differences between the two modes of administration) were intercorrelated *across* individuals. The resulting coefficients are listed in Table 8.

Results show that the individual retest stability in the verbal and non-verbal tests and the individual verbal and non-verbal validity correlate with one another. Moreover, the retest stability in verbal tests correlates with the validity of the verbal tests (.47 and .26 for control and fake good conditions, respectively), this also applies to the non-verbal tests (.54 and .57). Most importantly, the differences between the non-verbal and verbal stability are highly correlated with the differences between the validity coefficients of non-verbal and verbal tests. In other words, there is a strong tendency in the data that relatively large differences between the retest-stability of the non-verbal and verbal tests corresponded to larger differences in the coefficients of validity and *vice versa.*

Table 8: Correlations between individual coefficients of retest-stability and validity, separate for verbal and non-verbal tests. In the last two columns and rows, respectively, the correlations between the differences of stability and validity coefficients (non-verbal minus verbal) are shown. Each correlation is based on N = 95 subjects. Empty fields indicate non-significant correlations. All listed coefficients are highly significant. In the fields above the diagonal there are the correlations for the fake good condition, in the fields below, those for the control condition

	r_{tt} verbal	r_{tt} non-verbal	r_{tc} verbal	r_{tc} non-verbal	r_{tt} Difference	r_{tc} Difference
r_{tt} verbal	-	.58	.26	.33	.50	
r_{tt} non-verbal	.70	-	.33	.57	.42	.42
r_{tc} verbal	.47	.32	-	.63		
r_{tc} non-verbal	.40	.54	.60	-	.25	.68
r_{tt} Diff.		.62		.31	-	.26
r_{tc} Diff.		.39		.71	.45	-

These results imply that for some participants the non-verbal form provokes relatively reliable test responses in comparison with the verbal form and that the opposite is true for other participants. Further, the reliability of the non-verbal form is correlated with the validity of the non-verbal form. The findings with respect to the verbal form are analogous. However, it is also possible that the individual validity scores are only attenuated by the low individual reliability as might be expected from classical test theory.

Finally, it seemed of interest to examine the correlates of relatively high individual reliability and validity coefficients. For that reason, the coefficients of reliability and validity were correlated with the individual trait scores in the *PRF* and *NPQ*-scales. For the correlation between the test scores and the coefficients of *validity*, a spurious algebraic influence cannot be definitely excluded because the original individual test scores constitute not only one variable (X), but they are also included in the validity (as the mean standardized covariance between X and Y). To eliminate such an effect, the correlation between the coefficients of *stability* and the individual trait scores for the second ratings of the peers were used. The results are listed in Table 9.

Table 9: Correlations between individual coefficients of retest-reliability (for verbal and non-verbal items) and individual validity with trait-scores (in terms of peer-ratings) in PRF- and NPQ-scales

	r_{tt} verbal	r_{tt} non-verbal	r_{tc} verbal	r_{tc} non-verbal
PRF - Achievement	.30			.21
- Aggression	-.34		-.28	-.24
- Endurance	.39	.29	.25	.33
- Exhibition			-.20	
- Thrill Seeking				
- Nurturance	.25		.22	
- Social Recognition				
- Succorance				
- Understanding	.34	.35		.26
NPQ - Achievement				
- Aggression	-.40	-.26	-.37	-.38
- Endurance	.23	.23		
- Exhibition	-.29	-.24	-.34	-.28
- Thrill Seeking			-.30	-.21
- Nurturance	.22			.29
- Social Recognition	-.40	-.32	-.39	-.41
- Succorance		-.22		
- Understanding	.25	.31		.32

Empty fields indicate non-significant correlations. Coefficients $\geq .20$ $p < .05$; $\geq .25$ $p < .01$

As shown, the individual reliability in responding to verbal items is positively correlated with peer ratings on the trait dimensions *Endurance* (.39 *PRF* and .23 *NPQ*), *Nurturance* (.25 and .22, respectively) and *Understanding* (.34 and .25), negatively with *Aggression* (-.34 and -.40). The reliability in responding to non-verbal items correlates significantly only with two of the verbal scales (*PRF-Endurance* and *-Understanding*), but with six scales of the *NPQ*, thereby showing a very similar pattern to the reliability of responses to verbal items. Moreover the *NPQ-Succorance* scores in terms of peer ratings are correlated with individual reliability.

With regard to individual validity, the pattern of the correlations is by and large similar. To some degree, this may be due to some sort of spurious correlation, because the individual test scores which are used as *explanans* are also included in the validity coefficients as the *explanandum*. Therefore, the last mentioned results should be interpreted with caution until further empirical evidence with algebraic definitely independent measures are available.

Discussion

Our findings are easily summarized: in terms of reliability, validity, and the effects of social desirability, the verbal and non-verbal tests designed to measure Jackson's need constructs did not differ systematically. Both test forms showed about equal levels of internal consistency, stability and inter-rater agreement. They showed similar (and impressive) validities against peer criteria, and they showed similar (though small) correlations with a social desirability scale as well as similar effects in an experimental fake good manipulation.

These findings surprised us, as we had not expected the non-verbal test to work as well as Jackson's (1984) carefully developed *PRF* which includes more items for each scale. By inspecting the non-verbal test, we found some of the drawings ambiguous and difficult to understand. However, when we asked the study participants, informally, about their reactions to the two kinds of tests, we found an interesting divergence of opinion. Some participants agreed with our preconceptions, preferring the verbal test over the non-verbal one. However, others reported that they found the verbal items more ambiguous and difficult to understand. Future research may show which test is right for which individual.

Another interesting finding applies to social desirability effects in general. The statistical control of the valid variance leads only to a slight drop in the correlations between the *self*-ratings on the personality scales and the social desirability score. The magnitude of these semi-partial correlations is comparable to those between the *peer*-ratings on the personality scales and social desirability. This suggests that the social desirability score measures social desirable behavioral tendencies as well as the respective response style. Because social desirability scales are intended to measure only the response style, it must be concluded that they are "biased" by valid variance.

The observation that social desirability effects do not rule out validity applies not only to the correlational, but also to the experimental findings. Under the experimental fake good condition marked and well interpretable mean effects could be observed in comparison to the control group and Session 1. The scales with the highest degree of social desirability showed elevated means under fake good conditions. However, there was only a rather small decrease in validity and the amount of this decrease was not related to the social desirability of the

scales. This suggests that the faking effects are more or less additive, i.e. they cause only a general shift of the individual scores to the social desirable pole of the scale. Hence, these faking effects do not disturb valid variance substantially. This interpretation is supported by our finding that the correlations between the self-ratings from Session 1 and Session 2 under fake good conditions differed little from the retest-reliability found for the control group (see Table 2, fourth and last row). Only the *NPQ* scales *Succorance* and *Social Recognition* show substantially lower validity under fake good conditions ($r = .21$ vs. $r = .67$). But this finding is hardly interpretable and might be a retesting effect because the validity in Session 1 was only .37 in the control group and there was no mean effect on this scale.

So far, questionnaires might be a suitable personality measure and selection criterion despite their fakability - of course it should be kept in mind that in the present study the fakability was measured using a fake good instruction which presumably is different from faking under a real selection situation. Previous research focused mainly on demonstrating mean faking effects (e.g. Viswesvaran & Ones, 1999). However, the presence of (experimentally induced) faking effects does not preclude validity unless their size differs inter-individually. Nevertheless, valid comparisons are only possible between subjects who responded to the test under the same conditions. Hence, usual norm data are not appropriate under fake good condition.

References

1. Amelang, M. & Bartussek, D. (1970). Untersuchungen zur Validität einer neuen Lügen-Skala. [Studies on the validity of a new lie-scale.] Diagnostica, 16, 103-122.
2. Amelang, M. & Zielinski, W. (1997). Psychologische Diagnostik und Intervention. [Psychodiagnostics and Intervention.] Berlin: Springer.
3. Borkenau, P. & Ostendorf, F. (1987a). Untersuchungen zur faktoriellen Struktur retrospektiv geschätzter und on-line codierter Verhaltensfrequenzen: Eine Vergleichsstudie. [Examinations of the factorial structure of retrospectively estimated and on-line coded behavior frequencies: a comparative study.] Zeitschrift für Differentielle und Diagnostische Psychologie, 8, 259-274.
4. Borkenau, P. & Ostendorf, F. (1987b). Fact and fiction in implicit personality theory. Journal of Personality, 55, 415-443.
5. Costa, P. T. Jr. & McCrae, R. R. (1988). Personality in adulthood: A six-year longitudinal study of self-reports and spouse-ratings on the NEO personality inventory. Journal of Personality and Social Psychology, 54, 853-863.
6. Costa, P. T., McCrae, R. R. (1990). Personality disorders and the five-factor model of personality. Journal of Personality Disorders, 4(4), 362-371.
7. Gosling, S. D., John, O. P., Craik, K. H. & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. Journal of Personality and Social Psychology, 74(5), 1337-1349.
8. Jackson, D. N. (1967). Acquiescence response styles: Problems of identification and control. In I. A. Berg (Ed.), Response set in personality assessment (pp. 71-115). Chicago: Aldine.
9. Jackson, D. N. & Messick, S. (1967). Response Styles and the Assessment of Psychopathology. In D. N. Jackson & S. Messick (Eds.), Problems in Human Assessment (pp. 541-555). New York: McGraw-Hill.

10. Jackson, D. N. (1984). Personality Research Form manual. Port Huron, MI: Research Psychologists Press, Inc.
11. John, O. & Robins, R. W. (1994). Accuracy and bias in self-perception: Individual differences in self-enhancement and the role of narcissism. Journal of Personality and Social Psychology, 60, 206-219.
12. Mirels, H. L. (1982). The illusory nature of implicit personality theory. Logical and empirical considerations. Journal of Personality, 50, 203-222.
13. Murray, H. A. (1938). Explorations in personality. New York: Oxford University Press.
14. Paulhus, D. L. (1986). Self-Deception and impression management in test responses. In A. Angleitner & J. S. Wiggins (Eds.), Personality assessment via questionnaires (pp. 143-165). Berlin: Springer.
15. Paulhus, D., Fridhandler, B. & Hayes, S. (1997). Psychological defense: Contemporary theory and research. In R. Hogan & J. A. Johnson (Eds.), Handbook of personality psychology (pp. 543-579). San Diego: Academic Press.
16. Paunonen, S. V., Ashton, M. C. & Jackson, D. N. (2001). Nonverbal assessment of the Big Five Factors. European Journal of Personality, 15, 3-18.
17. Paunonen, S. V. & Jackson, D. N. (1979). Nonverbal trait inference. Journal of Personality and Social Psychology, 37(10), 1645-1659.
18. Paunonen, S. V. & Jackson, D. N. (1985). Idiographic measurement strategies for personality and prediction: Some unredeemed promissory notes. Psychological Review, 92, 486-511.
19. Paunonen, S. V. & Jackson, D. N. (1996). The Jackson Personality Inventory and the five-factor model of personality. Journal of Research in Personality, 30(1), 42-59.
20. Paunonen, S. V., Jackson, D. N. & Keinonen, M. (1990). The structured nonverbal assessment of personality. Journal of Personality, 58, 481-502.
21. Paunonen, S. V., Jackson, D. N., Trzebinski, J. & Forsterling, F. (1992). Personality structure across cultures: a multimethod evaluation. Journal of Personality and Social Psychology, 62, 447-456.
22. Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin, 86, 420-428.
23. Shweder, R. A. (1982). Fact and artefact in trait-perception: The systematic distortion hypothesis. In B. A. Maher (Ed.), Progress in experimental personality research, 1982, 11, 65-100.
24. Spinath, F. M. & Angleitner, A. (1995). Convergence of verbal and nonverbal personality assessment techniques. A German study using the NPQ. Bielefeld: Unpublished manuscript.
25. Stumpf, H., Angleitner, A., Wieck, T., Jackson, D. N. & Beloch-Till, H. (1985). German Personality Research Form (PRF). Göttingen: Hogrefe.
26. Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. Educational and Psychological Measurement, 59(2), 197-210.