“The impact of the test-taking context on responding to personality measures”

Masterthesis
Hede Höft
s0045802
University of Twente
February 2008

Introduction

For many decades, there has been an ongoing debate about the usefulness and validity of personality measures in selection contexts. While some authors believe that those instruments lack validity, others recommend their use as a selection instrument. There is plenty of literature available on this topic, but it is fairly controversial. Currently, a new discussion has evoked by the publication of an article by Morgeson et al. (2007a) on the occasion of a panel discussion at the 2004 SIOP conference in Chicago. Based on the literature review and the findings of the present study, this article will affiliate to this discussion.

However, validity is not the subject of the conducted study. In fact, the focus is on faking, which is also one of the main issues for arguments on personality measures in selection settings.

This article will first review the controversial existent literature on this topic, present the results of the present study, provide some possible explanations for the discrepancies in the literature and finally place the resulting insights in the current discussion between several authors, among which Ones et al. (2007) and Morgeson et al. (2007a, 2007b).

Validity of personality measures

The first important review about personality testing in organisations is provided by Guion and Gottier in 1965, who concluded that the validity of personality measures was too low to recommend its use as a basis for making employment decisions. Due to this and other rather pessimistic findings concerning this subject in that period, discussion about it faded for approximately the following 25 years. At the beginning of the 1990’s there was a revival of interest, due to the impact of personality in predicting job performance. This was mainly induced by two prominent meta-analyses (Barrick, & Mount, 1991; Tett, Jackson, & Rothstein, 1991) which stated that the lack of a well-accepted taxonomy of personality during the last couple of decades accounted for the discouraging findings concerning the relationships between particular personality constructs and performance criteria in different occupations. By that time, the five factor model ("Big Five") had emerged and evolved, which to its present claims to be the best paradigm for personality structure. Generally, researchers agree that there are five robust factors of personality (Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness) and the value of the model has widely been proved and is universally accepted. Therefore, Barrick and Mount (1991) examined the relationship of these five personality constructs to job performance measures for different occupations, rather than to focus on the overall validity of personality as previous researchers had done. Their results indicated that, as expected, Conscientiousness showed consistent relations with all job performance criteria for all occupational groups. Extraversion was a valid predictor for two occupations involving social interaction; managers and sales. Both Openness to Experience and Extraversion were valid predictors of the training proficiency criterion.

Similar to Barrick and Mount (1991) Tett et al. (1991) criticize the early work of Guion and Gottier (1965). They motivate the relevance of their review on the role of personality in job performance by the then current availability of more explicit conceptualizations of personality as related to work, the development of construct-oriented personality inventories, and the possibility of undertaking personality-oriented job analysis and advances in meta-analysis. They obtained a corrected estimate of the overall relation between personality and job performance of .24, but found that Conscientiousness showed lower validity than the other personality dimensions.

There are plenty other examples of studies (in a variety of occupational groups) which concluded that the Big Five are valid predictors of job performance. An example is the study conducted by Rust (1999), who found that all the Big Five traits have significant correlations with appropriate supervisors ratings. Another example is the study of Salgado and Rumbo

(1997), who examined financial services managers. They found that Neuroticism and Conscientiousness are correlated with job problem-solving ability, job motivation and global job performance. They also found that Extraversion, Openness and Agreeableness are correlated with global job performance. Detrick, Chibnall and Luebbert (2004) infer from their study that the use of the NEO PI-R (Costa, & McCrae, 1992) as a selection instrument for police officers appears promising. Ones and Viswesvaran (1996) believe that broadband personality variables, as the Big Five, have a great deal of potential for contributing to such theories as absenteeism, withdrawal behaviours, motivation, job satisfaction, and organizational commitment. As a result of his findings Salgado (2003) suggests that practitioners should use inventories based on the Five-Factor Model (FFM) in order to make personnel selection decisions. Hogan and Holland (2003) conclude that, as performance assessment moved from general to specific job criteria, all Big Five personality dimensions predicted relevant criterion variables more precisely with estimated true validities of .43, .35, .34, .36 and .34 respectively for Emotional Stability, Extraversion, Agreeableness, Conscientiousness and Openness to Experience.

"Faking" personality measures

Faking, which is also referred to as social desirability or impression management, is one of the main concerns with personality testing for selection purposes. It is likely that applicants in such a high stakes condition try to present themselves in a very good light, even if it is not the truth. It is also likely that they provide an exaggeration of some characteristics, in order to get hired. The question is not whether people *can* fake personality measures (i.e. when instructed to do so), since this fact has been proven plenty of times and researchers agree on it. Rather, the question is whether they *do* actually fake, to what extent and under which conditions.

Two authors who are notably engaged in investigating these aspects are Ones and Viswesvaran. In one of their meta-analyses (Ones, & Viswesvaran, 1996) they provide empirical evidence that (a) social desirability is not as pervasive a problem as has been anticipated by industrial-organizational psychologists, (b) social desirability is in fact related to real individual differences in Emotional Stability and Conscientiousness, and (c) social desirability does not function as a predictor, as a practically useful suppressor, or as a mediator variable for the criterion of job performance. In their opinion, removing the effects of social desirability from the Big Five dimensions of personality leaves the criterion-related validity of personality constructs for predicting job performance intact.

In a later meta-analysis Ones and Viswesvaran (1998) notice that their earlier work identified social desirability as the red herring in personality measurement. Their data from real-world job applicants confirm that criticizing personality scales because of potential response distortion by applicants is “making much ado about nothing”.

Similarly, using data from Project A, Hough et al. (1990) found that (a) personality scales are valid in predicting various on-the-job behaviours, (b) the criterion-related validities of personality scales are not destroyed even for individuals who are responding in an overly desirable manner and, (c) job applicant-like individuals’ responses to personality scales do not indicate distortion. On the basis of these findings Hough et al. (1990) concluded that personality scales could fruitfully be used in personnel selection and that social desirability did not moderate the personality-job performance relationships.

More recent studies come to the same conclusion. For instance, Kuncel and Borneman (2007), who present a new method for developing faking detection scales based on idiosyncratic item-response patterns in their article, used a within subjects design (first honest, later simulated application condition among students) and found evidence suggesting that faking is not, currently, a fatal problem.

Another team of researchers (Hogan, Barrett, & Hogan, 2007) who recently used a within subjects design (although this time participants were real job applicants who were rejected and reapplied for the same job six months later), found the same (for real-world selection settings).

Now, I will turn to those studies that *did* find significant differences between applicants' and incumbents' scores on personality measures and that *do* regard this as problematic and as a possible threat to the instruments' validities. However, there are also inconsistencies concerning the amount and the type of differences among the two groups of test-takers.

For example, in their meta-analysis of 29 studies comparing applicants with non-applicants, Birkeland et al. (2003) (in Weekley, Ployhart, & Harold, 2004) showed that across all jobs, applicants scored significantly higher on Emotional Stability, Conscientiousness, and

Openness.

Three years later, Birkeland et al. (2006) provided an extension of their meta-analysis, now including 33 studies comparing applicant and non-applicant personality scale scores. The earlier findings remained similar, with applicants scoring significantly higher on Emotional Stability (d=.44), Conscientiousness (d=.45), and Openness (d=.13). This time, they also reported applicants' significantly higher scores on Extraversion (d=.11).

The results of Weekley, Ployhart and Harold (2004) are in line with this, since their results suggest that mean score differences of applicants are three fourths of a standard deviation higher regarding Conscientiousness and half a standard deviation higher concerning Extraversion. In addition to previous findings, they also report higher applicant scores for Agreeableness (half a standard deviation). (However, they state that the criterion-related validities are not substantively affected by these issues; only slightly less in the applicant setting).

Despite a written warning, accompanying the directions for the applicants, that states that distorted self-descriptions would invalidate the respondents' test results, Zickar, Gibby and Robie (2004) found that applicants' mean scores were higher on all the Big Five scales. But their results also indicated that a high number of applicants score honestly and incumbents do fake, too.

Winkelspecht, Lewis and Thomas (2006) studied the same issue, but under experimental conditions instead of examining real-world applicants and incumbents. They showed that participants encouraged making a "most favourable impression" as a salesperson applicant score lower in Neuroticism and higher in Extraversion and Conscientiousness than participants encouraged responding "honestly". The authors conclude that response distortion may remain a serious threat to the use of personality test scores in selection.

Griffith, Chmielowski and Yoshita (2007) used a within-subjects design for their research and provided evidence that a significant number of applicants do fake personality based selection measures (30-50%) and that their score elevations resulted in significant rank ordering changes that impacted hiring decisions.

Another study evaluated the forced-choice format of items, considering its influence on faking behaviour (Heggestad, et. al, 2006). They found that scores based on multidimensional forced choice (MFC) response formats appear to provide normative information, but under faking conditions, they do not seem better at retaining the rank ordering of individuals than more traditional Likert formats. Similar to the earlier mentioned studies the authors conclude that in either case, faking distorts the rank order of the applicants, making it less likely that the best applicant will be hired.

Current discussion

The most current debate has been evoked by Morgeson et al.'s article (2007a) on the occasion of a panel discussion at the 2004 SIOP conference, where five former journal editors from *Personnel Psychology* and the *Journal of Applied Psychology* reconsidered the research on the use of personality tests in environments where important selection decisions are made. One of these participants, Michael Campion, who reviewed 112 articles on that topic, gives another demonstrative example of the inconsistency in literature. His statements most relevant for this article will be noted.

Concerning faking he found 14 studies comparing applicants to incumbents, of which seven reported higher applicant than non-applicant scores, four found that faking does occur, but not as much as expected and three found similar scores for both groups. Besides this, he concludes that directed faking studies show much greater effects of faking than studies of real applicants.

In regard of the question if faking affects criterion-related validity, Michael Campion's literature review came to similarly conflicting results. He found 18 studies, with eight finding that this is the case and ten finding that this is not the case.

The overall conclusions from Morgeson et al.'s article (2007a) are that (1) faking on self-report personality tests cannot be avoided and perhaps is not the issue; the issue is the very low validity of personality tests for predicting job performance, (2) using published self-report personality tests in selection contexts should be reconsidered and (3) personality constructs may have value for employee selection, but future research should focus on finding alternatives to self-report personality measures.

Ones et al. (2007) responded to this article and concluded (among others) on the basis of several meta-analyses that (1) personality variables, as measured by self-reports, have substantial validities, (2) self-reports of personality, in large applicant samples and actual selection settings, have yielded substantial validities even for externally obtained and objective criteria and (3) faking does not ruin the criterion-related or construct validity of personality scores in applied settings.

Thereupon Morgeson et al. (2007b) provided another article to address some of the just mentioned points made by Ones et al. (2007) and the negative appraisal of other authors. In that paper Morgeson et al.'s (2007b) main criticism about personality tests and their use for personnel selection remains their low validities, and they argue against Ones et al's conviction that the types of corrections that have been applied to the observed validities of personality tests have not changed and that these are responsible for the optimism about the usefulness of personality tests for personnel selection.

Besides the validity of personality measures in selection settings, the two main parties in the above mentioned argument, Ones et al. and Morgeson et al., also disagree on the faking issue. While the first thinks that only few applicants actually intend to fake in real employment situations, the latter state that faking on self-report personality tests should be expected and cannot be avoided.

To shed some light on this aspect, the present research will focus on faking personality measures in different situations and the research question is formulated as follows:

Are there any significant differences between the personality scale scores obtained in a selection setting and the scores obtained in a developmental context?

Method

Sample

The data were provided by a Dutch company that implements solutions in the field of performance management, leadership, competence development and selection services. They developed several personality measures which are used in different (selection- and developmental) contexts, totalling roughly 51,000 data-sets of "real world" administrations collected between May 2004 and October 2007. The large data set was explored and to keep as many variables as possible constant, an organisation was identified that used both instruments for the same job functions in both selection and developmental contexts. This resulted in a data set with 922 administrations that was used for analysis.

The set refers to an organisation which provides ICT services and consulting. Of the respondents who completed the questionnaires 17% were female and 83% male. Concerning the job functions, 42 were Analysts, 102 Business Consultants, 140 Business Intelligence Consultants, 21 Business Unit Managers, 53 ICT Consultants, 63 Functional ES Consultants, 16 Management Consultants, 19 Developers, 58 Project Leaders, 20 Project Managers, 18 Secretaries, 138 Software Engineers, 9 Staff Employees, 23 System Engineers, 54 Technical Specialists and 153 Trainees. The respondents reported an average of 6.3 years spent in the workforce (SD=6.89). Furthermore, the entire sample was educated beyond high school and its vast majority had an educational level of graduate or undergraduate (39% and 53%, respectively). 663 tests were administered in a selection context and 259 in a developmental context.

When comparing the two contextual groups (selection and development) no notable differences concerning the distributions of sex, function and educational level are discovered. Only in regard of work experience there is some variance, since about 39% of the respondents in the selection context have spent no or less than one year in the workforce, while this only accounts for 3% in the developmental context.

Instruments

Workplace Big Five (WB5; Schakel, Smid, & Jaganjac, 2007a). The WB5 is an online instrument that gives global insight into a candidate's personality and his/her disposition for developing certain (43) competencies. The 144 items refer to behaviour that is relevant for the work situation, based on the Big Five personality model with its basic characteristics Emotional Stability, Extraversion, Openness, Agreeableness, and

Conscientiousness and its 24 underlying facets. Scoring is based on a 5-point rating scale from 1 (strongly disagree) to 5 (strongly agree). Its administration takes about 20 minutes.

The alpha reliability coefficients for the 24 facets range between 0.66-0.82 with a mean of 0.75. The internal consistency estimates for the scores on the main Big Five factors are .87, .91, .90, .86 and .93 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively. The correlations between the WB5 and the NEO-PI-R are

.73, .71, .31, .42, and .68 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively (estimates corrected for attenuation: .81, .78, .34, .48 and .77). The predictive validities for the 43 competencies vary between .25 and .65.

Connector P (Conn P; Schakel, Smid, & Jaganjac, 2007b). The Connector P is also an online personality measure and could be characterized as the short form of the WB5, since it is constituted of half of its 144 items. Additionally, it contains 10 items measuring self-image, indicating the way in which the candidate positions himself in regard to others. It uses the same response options.

Given that Conn P is a “light” version of Workplace Big Five, the psychometric properties are the same for both instruments, except from the effect of shortening test length. The internal consistency coefficients are .78, .85, .80, .75, .86 for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness, respectively and .65 for Self-Image.

The internal consistencies for the competency estimates vary between .63 and .88.

Procedure and Design

The conducted study has a non-experimental, cross-sectional, between-subjects design. As already mentioned above, an appropriate data-set to be analyzed was identified, resulting in 922 test scores that were made anonymous. To allow for comparing the same constructs measured by both instruments, new variables were created by aggregating the mean scores on the identical items in the two questionnaires. This amounted to 29 variables; 24 facet mean scores and 5 factor mean scores.

Results

Mean Differences

First of all, the entire mean scores on the Big Five factors and its facets were compared in respect of their test-taking context.

Several significant differences between the scores in selection and developmental context were found. At a significance level of p<0.01, discrepancies were found for the scores on 14 of the 24 facets and three of the factors and for four more facets at level p<0.05. The effect sizes range between .20 and .53 with a mean of .31 for the facets and .33 for the factors. The scores in the selection context are lower on two Neuroticism facets and one Agreeableness facet. The rest of the mean scores are higher. Also, the mean scores on the factors Extraversion, Openness and Conscientiousness are higher in the applicant pool, with the biggest effect size for the last-mentioned (more than ½ standard deviation). For the exact significant effect sizes per facet and factor see Table 1 and 2.

Next, only the mean scores of one of the instruments were regarded, in this case the WB5, since the Conn P was not used in both contexts. Again, the scores between selection and developmental context were compared. Here, significant differences at p<.0.01 were found for 11 of the 24 facets and for 3 factors. At p<0.05 significant differences were found for five more facets and one factor. The significant effect sizes are even higher with a range of .33 to .73 and a mean of .46 for the facets and .54 for the factors. The results show that on average, applicants score significantly lower on Neuroticism and higher on Extraversion, Openness and Conscientiousness with, again, the largest differences concerning the last-mentioned (almost ¾ standard deviation). For the exact significant effect sizes per facet and factor see Table 1 and 2.

After that, the data were even more specified by choosing only the scores of persons with the same job description. This led to the selection of only Software Engineers (both applicants and incumbents) in the ICT Service Organisation. The results identified seven significant differences at level $p<0.01$ and six at level $p<0.05$ regarding the facets. Concerning the factors, the only significant result was found for Conscientiousness, but with a huge effect size of almost one entire standard deviation. The significant effect sizes for the differences between the mean facet scores vary between .40 and .88 with a mean of .56. For the exact significant effect sizes per facet and factor see Tables 3 and 4.

Subsequently, persons with another concurrent job function were chosen, resulting in the selection of the data of all Business Consultants in the ICT Service Organisation. This time, not a single significant difference was found between the applicants' and incumbents' scores on the personality measure. Therefore, the mean scores between the two occupational groups were compared, once only the tests taken in a selection context were regarded and once only those tests taken in a developmental context. The results showed effect sizes above .20 for all but three facets and for all the factors except from Agreeableness. Seven of these standardized mean score differences are significant at level $p<0.01$ and other seven at level $p<0.05$.

The significant effect size for the facets averages .68. Most noticeable are the Extraversion facets Sociability, Taking Charge and Directness, on which the Business Consultants score, on average, .73, .77 and .95 standard deviations higher than the Software Engineers. Also the differences between the scores on the Conscientiousness facets Organisation and Drive are quite sizable, namely .75 and .76 standard deviations higher for the Business Consultants.

The same analysis was repeated, but this time only the data from selection contexts were included. Then, some of the effects disappeared, and “only” 14 of the 24 facets had an effect size above .20 with an average of .52. Of these, ten facet scores were significant at level p<0.01 and two at level p<0.05. While the effects specifically concerning the facets belonging to Extraversion and Openness remained, either disappeared or changed those concerning the Conscientiousness facets. Whereas in the developmental context the Software Engineers scored lower on all five Conscientiousness facets (of which three significantly) and the factor itself, in the selection context their scores were only significantly lower on one C-facet and the effect on the factor Conscientiousness even disappeared almost entirely. The significant effect sizes regarding Neuroticism, Extraversion and Openness remained.

Additionally, to explore if there are any gender differences concerning responding to personality measures, men's and women's scores on the five factors were also compared. In the whole sample (N=922) men scored significantly lower than women on Neuroticism (d=.27, p<0.01), Extraversion (d=.35, p<0.01) and Agreeableness (.19, p<0.05). Men's scores were higher on Openness (d=.28, p<0.01).

When comparing the women's scores across the two contexts, results show that in selection, their scores were significantly higher (p<0.01) on Openness and Conscientiousness than in a developmental context (d=.45 and .50).

When men took the personality test in a selection setting, they also scored significantly higher on Openness and Conscientiousness and additionally on Extraversion (d=.32, d=.54 with p<0.01 and .21 with p<0.05, respectively). Their effect size for Openness was smaller than the women's, but for Conscientiousness it was a little bigger. Table 5 gives an overview of the exact effect sizes for each condition the two groups were compared in.

Discussion

The research question, if there are any significant differences between the personality scale scores in a selection context and the scores obtained in a developmental context, is clearly affirmed.

Especially in regard of the factor Conscientiousness, the presumption that applicants score significantly higher than non-applicants, was confirmed. Depending on which variables were kept constant, the effect sizes for this factor ranged between .53 and .94, which is notably high. These effects were expected, since such a high stakes condition as application makes the test-takers present themselves in the most favourable light. Besides this, Conscientiousness seems to be the most important factor concerning job performance through all kinds of branches, functions and positions. Also, applicants' higher scores on Extraversion and Openness were expected, since most people regard these as desirable characteristics concerning work life. The effect sizes were somewhat smaller than those in respect of Conscientiousness, but still quite meaningful with significant values between .21 and .58.

The present study has found clear evidence for the fact that applicants tend to inflate their scores on several desirable personality characteristics and supports the findings of other authors, among which Birkeland, et al. (2006) and Weekley, Ployhart and Harold (2004). Nevertheless, many researchers found different or opposite results and this study was conducted to shed some light on these discrepancies that still persist, even after the development of better research methods, the definition of a clear and universally accepted personality theory (FFM) and other empirical improvements during the last decades. But although the literature is so diverse about faking personality measures in a selection context, no reasons are assigned for the partially opposite findings. So what accounts for these discrepancies?

By searching the literature and reading a great variety of articles concerning this issue, it becomes clear, that the research conditions to some extent differ extremely and the assumption suggests itself, that these differences provide a possible explanation. Differences in the samples, (i.e. concerning its size and the type or group of respondents) and the way of detecting faking (either through social desirability scales or calculating the standardized mean differences) are only two examples of important issues that can influence the outcomes of a study.

Moreover, one must keep in mind that there are several possible moderator variables that can influence the results of research on this topic, among which job description/position and branch of trade/industry sector, instrument and response format, nature of study design (within-subject, between-subject, real-world, induced-faking, etc.). In the next passage these issues will be addressed in succession.

The present study provides an example of the importance of considering all of the above mentioned variables and interpreting results carefully. The results are seen as meaningful, since the research sample was selected carefully with regard to choosing only one organisation and a quite equal distribution in both contextual groups concerning demographic variables and job function. The last mentioned variable is probably most important in influencing personality scale scores, more than age, sex, and years spent in workforce or similar. The results of the present study indicate this, since there are severe differences between Software Engineers and Business Consultants, for instance. Software Engineers appear to inflate their scores immensely on the factor Conscientiousness (almost one standard deviation), while they apparently see no reason to do the same with Extraversion or Openness, since these characteristics do not seem to play an important role in their job description. On the other hand, Business Consultants in general have high scores on all of these three factors, so there is no reason and no real option anyways to significantly inflate/overstate those in order to get hired for a job. But many other studies, especially the meta-analyses, do not explicitly differentiate between job functions, which might have an impact on the results.

Some authors differentiated at least to some degree between occupational groups, as Birkeland, et al.(2006) who compared Sales vs. Non-Sales and Management vs. Non-Management occupational groups, which might be one of the reasons for the similar results that were found in regard to the present study.

Another variable that can influence research results is type of instrument. For the literature review in this article only studies applying the five-factor model were used, but even while intending to measure the same constructs, instruments can still vary immensely. For instance, in Hogan, Barrett and Hogan's study (2007) applicants completed the Hogan Personality Inventory (HPI; Hogan, & Hogan, 1995), a 206-item, true-false inventory of normal personality designed to predict occupational performance, containing seven primary scales that align with the five-factor model of personality. But the underlying facets and also the response format are not the same as in the presently used personality measure, which disallows comparisons. Besides this, it was a within-subjects design with applicants that were rejected the first time and then reapplied for the same job 5 months later. Based on the change scores the authors concluded that faking on personality measures is not a significant problem, but the test-taking conditions in this case were very specific and also the motivation to fake or not to fake differed.

As a matter of course the study design influences the results. The main distinction that must be made concerning faking personality tests is whether it is a real-world or an induced-faking study. One example that demonstrates the possible impact is provided by Birkeland et al.(2006) who hypothesized smaller effect sizes for their study than in one of Viswesvaran and Ones' studies (1999), since they conducted a 'real-world' study with authentic applicants in contrast to comparing induced 'fake-good' vs. honest responses with an experimental character, and found evidence for their hypothesis.

The results of this study do not provide evidence for notable sex-differences in social desirable responding.

However, it is interesting to note that, although not significantly, both sexes lower their scores on Neuroticism when in a selection setting, but women also lower their scores on Extraversion and Agreeableness, while men inflate those. This results in about the same scores, with no significant differences left when comparing the two groups only in selection context. Based on this finding one could infer that members of both gender have a very similar opinion about what kind of personality is desirable for a certain kind of job function and they present themselves accordingly, hoping to please the employer’s demands.

A similar conclusion can be drawn from the results concerning the occupations Software Engineers and Business Consultants. It seems that the applicant is aware of the demands of the desired job and that he/she is able to and does in fact adjust his/her personality test scores to the extent he/she thinks he/she will perfectly fit the job in the eyes of the potential employer. Recruiters and employers should be aware of this fact.

Based on the literature review and the validity studies accompanying the Workplace Big Five it is concluded that, in support of Ones et al. (2007), the Big Five are valid predictors of job performance and that personality measures based on them have substantial criterion-related validity and therefore should be used in organizational decision making, including personnel selection. Of course, and even Morgeson et al. (2007a, 2007b) admit this, some instruments are better than others. I regard the WB5 and the Conn P as examples of good instruments for use in selection. These instruments are very carefully and methodologically sound developed and, most important, refer explicitly to personality at work and not in general. Morgeson et al. (2007b) recommend adding “at work” to the items to achieve better empirical results. The WB5 and Conn P meet this demand by explicitly referring to work situations. 43 competencies were identified and evidently the personality profile measured can predict the scores on them (correlations range between .25 and .65), corroborating the validity of the Big Five personality model in predicting work performance.

Furthermore, Morgeson et al. (2007b) note that the validity of self-report personality measures is likely to be greater when they are used in combination with cognitive ability tests. I fully agree with this and wonder if, in practice, there is any organisation or person in charge at all that would use a personality test by itself to make a selection decision. Rather, often some sort of cognitive ability test is used in addition, and in approximately all cases an interview is associated with the personality test, alone to give the respondent feedback on his/her test results. At least, this would ethically be correct. The WB5 and the Conn P administration provide both. The interview also serves as an add-on to validity and solves, at least for the main part, the only problem I consider with personality tests in selection settings: faking. As described above, the present study found notable evidence for respondents tending to inflate their scores, especially on Conscientiousness, if they are in an application situation. But when people are trained in administering the WB5 and Conn P, they also learn what questions to ask to check the validity of the respondents' answers on the test and thereby compensating the eventually inflated scores.

Compared to previous studies, the present research incorporates the important aspects that are essential for achieving sound results on this topic. That is to say, the sample consisted of real world applicants and incumbents and had a reasonable size. People in the same organisation and with the same job descriptions were compared. Besides, the instruments used are valid, based on the FFM, and work-related. This study should be replicated in other organisations and with different occupations. To elaborate research on faking and to even more specify results, one improvement could be implemented by further research. That is using a within-subject design instead of a between-subject design to eliminate irrelevant variables that might have an influence on the results.

Main Conclusions

1. There are carefully developed and sound self-report personality measures available that are valid instruments which, in combination with other diagnostic instruments (i.e. cognitive ability test), should be used for selection.
2. Faking does occur in selection settings and the effects are the highest for Conscientiousness. Which scores are inflated (besides Conscientiousness) depends especially on the job description/occupation. One can argue whether the inflation is a

constant shift or if it is even desirable if people fake, since it could be seen as an expression of ability to adapt and meet expectations. But independent of one’s opinion about faking, one can prevent hiring the wrong (intentionally faking) people, by providing a good feedback interview and ask clarifying questions.

References

Barrick, M. R., Mount, M. K. (1991). The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology, 44*, 1-26

Birkeland, S., Manson, T. M.., Kisamore, J. L., Brannick, M. T., Smith, M. A. (2006). A Meta-Analytic Investigation of Job Applicant Faking on Personality Measures. *International Journal of Selection and Assessment, 14*(4), 317-335

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

Detrick, P., Chibnall, J. T., Luebbert, M. C. (2004). The Revised NEO Personality Inventory As Predictor of Police Academy Performance. *Criminal Justice and Behavior 31*(6), 676-694

Griffith, R. L., Chmielowski, T., Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behaviour. *Personnel Review, 36*(3), 341-355

Guion, R. M., Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18,* 135-164

Heggestad, E. D., Morrison, M., Reeve, C. L., McCloy, R. A. (2006). Forced-Choice Assessments of Personality for Selection: Evaluating Issues of Normative Assessment and Faking Resistance. *Journal of Applied Psychology, 91*(1), 9–24

Hesse, J. & Schrader, H. C. (2002). Assessment Center: das härteste Personalauswahlverfahren bestehen. Eichborn Verlag, Frankfurt

Hogan, J., Barrett, P. & Hogan, R. (2007). Personality Measurement, Faking, and Employment Selection. *Journal of Applied Psychology, 92*(5), 1270-1285

Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory manual.* Tulsa, OK: Hogan Assessment Systems.

Hogan, J. & Holland, B. (2003). Using Theory to Evaluate Personality and Job-Performance Relations: A Socioanalytic Perspective. *Journal of Applied Psychology, 88*(1), 100-112

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581-595

Kuncel, N. R.& Borneman, M. J. (2007). Toward a New Method of Detecting Deliberately Faked Personality Tests: The use of idiosyncratic item responses. *International Journal of Selection and Assessment, 15*(2), 220-231

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., Schmitt, N. (2007a). Reconsidering the use of personality test in personnel selection contexts. *Personnel Psychology, ?,* 683-?

Morgeson, F. P., Campion, M.A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., 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

Ones, D. S., Dilchert, S., Viswesvaran, C., Judge, T. A. (2007). In Support of Personality Assessment in Organizational Settings. *Personnel Psychology, 60,* 995-1027

Ones, D. S., Viswesvaran, C. (1998). The Effects of Social Desirability and Faking on Personality and Integrity Assessment for Personnel Selection. *Human Performance, 11* (2/3), 245-269

Ones, D. S., Viswesvaran, C., Reiss, A. D. (1996). Role of Social Desirability in Personality

Testing for Personnel Selection: The Red Herring. *Journal of Applied Psychology, 81* (6), 660-679

Püttjer, C. & Schnierda, U. (2004). Assessment Center Training für Führungskräfte; die wichtigsten Übungen – die besten Lösungen. Campus Verlag, Frankfurt/New York

Rust, J. (1999). Discriminant Validity of the Big Five Personality Traits in Employment Settings. *Social Behavior and Personality 27*(1), 99-108

Salgado, J. G. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*, 323-346

Salgado, J. F. & Rumbo, A. (1997). Personality and Job Performance in Financial Services Managers. *Personality and Job Performance, 5*(2), 91-100

Schakel, L., Smid, N. G., & Jaganjac, A. (2007a). *Workplace Big Five professional manual.* Utrecht, The Netherlands: PiCompany B.V.

Schakel, L., Smid, N. G., & Jaganjac, A. (2007b). *Workplace Big Five professional manual.* Utrecht, The Netherlands: PiCompany B.V.

Schwertfeger, B. (2005). Chancen erkannt. *Capital, 20,* 102-106

Tett, R. P., Jackson, D.N., Rothstein, M. (1991). Personality Measures as Predictors of Job Performance: A Meta-Analytic Review. *Personnel Psychology, 44,* 703-742

Weekley, J. A., Ployhart, R. E. & Harold, C. M. (2004). Personality and Situational Judgement Tests Across Applicant and Incumbent Settings: An Examination of Validity, Measurement, and Subgroup Differences. *Human Performance, 17*(4), 433-461

Winkelspecht, C., Lewis, P., Thomas, A. (2006). Potential effects of faking on the NEO-PI-R: Willingness and ability to fake changes who gets hired in simulated selection decisions. *Journal of Business and Psychology, 21*(2), 243-259

Zickar, M. J., Gibby, R. E., Robie, C. (2004). Uncovering Faking Samples in Applicant, Incumbent, and Experimental Data Sets: An Implication of Mixed-Model Item Response Theory. *Organizational Research Methods, 7*(2), 168-190

Table 1

Standardized Mean Differences between Applicant and Incumbent Facet Scores

Facet	Applicant - Incumbent	Applicant – Incumbent, only WB5
N1 Sensitiveness	-.245**	--
N2 Intensity	--	--
N3 Interpretation	-.300**	-.398*
N4 Rebound Time	.317**	--
N5 Retiecence	--	-.413**
E1 Enthusiasm	--	.520**
E2 Sociability	.250**	.530**
E3 Energy Mode	.252**	.458**
E4 Taking Charge	--	.560**
E5 Directness	--	--
O1 Imagination	.297**	.442**
O2 Complexity	--	.448**
O3 Change	--	.326*
O4 Autonomy	.242**	--
A1 Service		.376*
A2 Agreement	-.207**	--
A3 Deference	--	--
A4 Trust in Others	--	--
A5 Tact	.391**	.483**
C1 Perfectionism	.500**	.366*
C2 Organisation	.327**	.560**
C3 Drive	--	.600**
C4 Concentration	.468**	.391*
C5 Methodicalness	.321**	.590**

Note. The effect sizes in each cell represent the mean difference (first group minus the second group) divided by the pooled standard deviation. Bold values are significantly different.
**= p<0.01 and *=p<0.05

Table 2

Standardized Mean Differences between Applicant and Incumbent Factor Scores

Factor	Applicant – Incumbent	Applicant – Incumbent, only WB5
Neuroticism	--	**-.398***
Extraversion	**.206****	**.580****
Openness	**.261****	**.447****
Accommodation	--	--
Conscientiousness	**.531****	**.730****

Note. The effect sizes in each cell represent the mean difference (first group minus the second group) divided by the pooled standard deviation. Bold values are significantly different.
**= p<0.01 and *=p<0.05

Table 3

Standardized Mean Differences for Software Engineer and Business Consultant Facet Scores

Facet	Software Engineers Applicant – Incumbent	Incumbents Software Engineers – Business Consultants	Applicants Software Engineers – Business Consultants
N1 Sensitiveness	**-.464***	--	--
N2 Intensity	**-.403***	.380	--
N3 Interpretation	**-.439***	.466	**.347***
N4 Rebound Time	**.458***	--	**.583****
N5 Retiecence	--	**.626***	**.714****
E1 Enthusiasm	--	**-.588***	--
E2 Sociability	--	**-.728****	**-.628****
E3 Energy Mode	**.489***	**-.774****	**-.734****
E4 Taking Charge	--	**-.953****	**-.907****
E5 Directness	--	-.272	**-.481****
O1 Imagination	**.552****	-.457	-.202
O2 Complexity	--	**-.561***	**-.439****
O3 Change	--	**-.535***	**-.800****
O4 Autonomy	--	-.487	**-.341***
A1 Service	--	-.486	--
A2 Agreement	--	.488	**.452****
A3 Deference	**-.519****	**.608***	--
A4 Trust in Others	--	--	--
A5 Tact	**.685****	-.455	--
C1 Perfectionism	**.562****	-.216	--
C2 Organisation	**.883****	**-.752****	--
C3 Drive	**.412***	**-.760****	--
C4 Concentration	**.832****	-.415	.216
C5 Methodicalness	**.597****	**-.571***	--

Note. The effect sizes in each cell represent the mean difference (first group minus the second group) divided by the pooled standard deviation. Bold values are significantly different.
**= p<0.01 and *=p<0.05

Table 4

Standardized Mean Differences for Software Engineer and Business Consultant Factor Scores

Facet	Software Engineers Applicant – Incumbent	Incumbents Software Engineers – Business Consultants	Applicants Software Engineers – Business Consultants
Neuroticism	--	.357	**.567****
Extraversion	--	**-.954****	**-.834****
Openness	--	**-.604***	**-.582****
Accommodation	--	--	.210
Conscientiousness	**.944****	**-.766****	--

Note. The effect sizes in each cell represent the mean difference (first group minus the second group) divided by the pooled standard deviation. Bold values are significantly different.
**= p<0.01 and *=p<0.05

Table 5

Standardized Mean Differences for Men and Women Factor Scores

Facet	Men- Women (general)	Men- Women (Selection)	Men Selection- Development	Women Selection- Development
Neuroticism	**-.266****	**-.237***	-.103	-.226
Extraversion	**-.347****	-.199	**.318****	-.246
Openness	**.267****	.190	**.207***	**.447****
Accommodation	**-.190***	-.136	.039	-.134
Conscientiousness	-.086	-.101	**.543****	**.499****

Note. The effect sizes in each cell represent the mean difference (first group minus the second group) divided by the pooled standard deviation. Bold values are significantly different.
**= p<0.01 and *=p<0.05