A third generation personality test

Lennart Sjöberg
Center for Risk Research and Center for Media and Economic Psychology
Marketing and Strategy Department
Stockholm School of Economics

No. 2010:3

Revised 2012

Contents

Abstract

The development of personality testing (self-report) in the workplace has undergone three phases. The first generation of tests, such as Cattell's 16PF and OPQ, was characterized by complex systems for the description of the personality. These systems were simplified in part by the following generation of tests, which was based on the five-factor model but that model was simple only at the over-arching level. Beneath the five main factors were a large number of ancillary factors, usually 30-40 in number. No tests of the first and second generation could effectively handle the problem of impression management, nor did they take into account the effects of mood on the test responses. These and a number of other problems were solved to a great extent in the *UPP* test, which therefore is proposed to represent a third generation of personality tests. The test focuses on "narrow" and work-relevant traits, and includes aggregated variables with the same focus, including two variables especially fitted to the requirement of any given application, an effective and validated method for correction for impression management, extensive treatment of quality of data from each tested person to yield a "warning signal" when results should not be trusted, as well as measurement of current mood at the time of testing which can give another "warning signal." Measurement of attitude toward the test ("face validity") is included, and measures of work-related attitudes, which are of value in themselves but can also be used as proxy criteria, greatly facilitating validation work. The true score variance of the UPP test scales not accounted for by the Big Five is estimated to 49 %, implying that the test provides an abundance of information beyond the Big Five. Validity against external criteria has been found to be at the 0.5 level.

First generation: 1920-1990

Self-report tests of personality go back to the First World War. Until the 1960s the American tradition was represented by factor tests such as Cattell's [9] 16PF - 16 factors. Eysenck's much simpler MPI [23], developed in the UK, had only three factors. The American tradition, however, came to dominate. Complex tests were most common. However, Cattell, creator of the 16PF, based his test on factor analysis of small data sets. The results were hard to replicate [13]. The fact that the first generation of personality tests resulted in very complicated test structures was probably due to the fact that there was no access to effective statistical and psychometric methodologies. The validity was questionable. As an example, early Swedish work on the 16PF showed that it failed to predict manager performance [61].

Some traditional British tests are often used in Sweden: PPA [37] and OPQ [24]. They are based on older and, in the case of the OPQ, quite complex models. OPQ32r measures 32 personality variables with relatively few data, a total of 104 blocks with three statements in each block. The reliability of some of the 32 scales is somewhat low. A meta-analysis of British research on the OPQ [64] showed that the test had a validity of around 0.2, which is significantly worse than most other personality tests that usually lie at 0.3. In the latest development of the OPQ, it has been shown that the test has a five factor ("Big Five") structure [6].

A special report on the validity of OPQ for sales work showed a mean validity around 0.1 [71]. This is comparable to what is reached in another specialized trend in testing: integrity tests [96].

The practical application of tests is changing only slowly, if at all. The most widely used test in Sweden, PPA, is an adjective list with roots in the 1920s [47]. It possibly measures at least one important dimension [45], but far from all of potential importance in industrial psychology applications. Another old test, MBTI, is based on Jung's personality theory from the early 1920s [41], and has no explanatory power with regard to job performance. See a critical discussion of MBTI elsewhere [75] or Paul's lucid review [58].

Most of the first-generation tests were flawed in that they resulted in an extremely complex picture of personality. They placed great demands on the experience and intuitive ability of the psychologists who used them. Validation research showed that the validity was low, at the most around 0.3. These tests are marketed as providing material for more or less speculative interpretations, not as prognostic instruments. It is often explicitly stated that there are no univocal relationships between test scale results and criteria - they must apparently be inferred by the user, or provided in narrative descriptive texts with unclear foundations. However, research has not been able to confirm the existence of non-linear relationships between test scores and criteria [97].

Validity around 0.3 is a finding that Mischel drew important conclusions from in his classic book from 1968 [52], but already in 1921 the Allport brothers reported data at the same low level [1]. They speculated then that it would be possible to obtain much stronger results with better tests. It has since become apparent that it was very hard to reach that goal.

One of the basic problems of the first generation of personality tests was the concern with prediction. This is in itself a worthy goal, but it is frequently unrealistic. Even if we have very effective and relevant measures of the personality dimensions, there can be no guarantee that we can make highly efficient predictions of how a person will succeed in a certain environment or with a specific task. There are many aspects other than his or her personality who came into play. We do not even know how much of behavior, or job success, which, in principle, can be predicted based on individual factors - the question is rarely discussed.

Second generation: 1990 - 2010

Around 1990, the notion of the “Big Five” was suggested [49]. It seemed to be a very attractive notion to apply just five factors to describe personality. American tests such as NEO-PI-R [14] and HPI [35] are based on this model. The international database IPIP has released test items, which may be translated and used to measure the Big Five factors. These are:

- Extraversion
- Emotional stability
- Conscientiousness
- Openness
- Agreeableness

The Big Five tests are the second generation of personality tests. They are less complex than the first-generation tests in the sense that they focus on relatively few over-arching dimensions. However, complexity is still left in the form of subscales, or facets. In the NEO-PI-R, there are 32 such subscales and complexity is therefore significantly higher than in similar tests of the first generation such as the 16PF, not to mention the structurally simpler MBTI and PPA.

The second-generation tests, the Big Five tests, created a paradoxical difficulty: they are both too general and too detailed. On the overall Big Five level the tests are weak in relation to the relevant criteria, but the detailed level, such as "facets" of the NEO-PI-R, gives a number of scales, which are hard to manage effectively. Thirty or forty scales are probably too much to handle cognitively, and doubts arise if all of these scales can be measured reliably with a test taking about 30-40 minutes. The consequences of complexity are further discussed later in this paper.

Second-generation Big Five tests have had a strong impact. It is often claimed that new tests should be evaluated against the Big Five to see if they add some value to the prognostic power of the tests. Extensive studies have been made of the Big Five dimensions regarding their practical value in the workplace. The results have been disappointing, however [53; 54]. Big Five dimensions have not produced an improved predictive power [80].

A basic problem is connected with the broad, over-arching nature of the FFM dimensions. It is becoming increasingly clear that more narrow scales are needed in order to predict important job performance criteria [4; 44; 59; 92].

Second generation tests also did not solve the problem of impression management (IM) in high-stakes testing [84]. The large number of current publications about faking on personality tests testifies to the importance of the topic. The database PsychINFO lists 1358 entries for

the search term "faking", the oldest one from 1922. The time trend of publications dealing with the topic is illustrated in Fig .1.

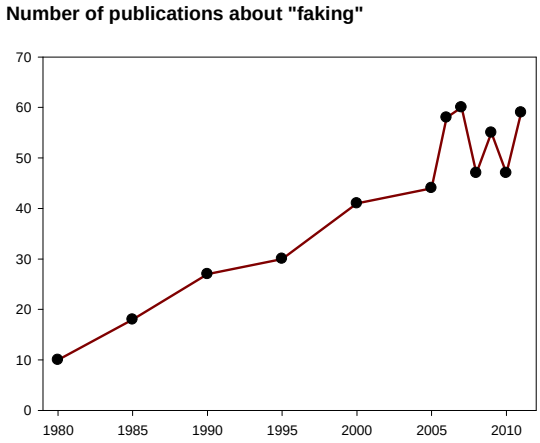

Figure 1. Number of publications about "faking" in PsychINFO since 1980.

There is a clear and steady increase, and at present a stable level of interest, producing some 50 publications each year. Apparently, the problem exists.

Mood states have been shown to affect cognitive process such as expectations [73; 87]. Few studies have investigated mood and responses to personality tests; for an exception see [48]. There is good reason to believe, however, that unusual mood states affect how people see themselves and hence how they respond to a personality test. Test results could be quite misleading if the is factor is not taken into account, but this is not done in standard personality tests.

Another factor seldom attended to is how the tested persons regard to the tests, or "face validity". It is, of course, not likely that tested persons can make a highly accurate judgment validity. E.g., it was found that projective tests were believed to be especially valid [69], a belief known to be false [27; 99]. Reactions to a selection program depend on perception of the quality and fairness of the tests being used [28]. Tests should include means of assessing these perceptions, both in order to monitor how selection programs are perceived, and to investigate individual cases of extreme reactions.

Some tested persons respond in a haphazard and little motivated manner. It is important to assess the quality of data for each individual. Interpretation of the test results is compromised if the data are of low quality. Such problems may also be due to lack of proper understanding of instructions or test items, or highly idiosyncratic thinking. Data quality should be assessed as a matter of routine in personality testing.

Third generation

The first and second generations of personality tests faced some important problems. Summing up, the main problems were:

1. Low explanatory power with regard to job performance and training results
2. No credible way of correcting results for IM

3. Lack of ways of measuring data quality, mood state and face validity (attitude toward the test)

I shall now discuss these problems and start with impression management. My examples come from the UPP-test[1].

The *UPP* test

UPP is a personality test intended for applications in industrial and organization settings. The test consists of 348 items. Most are self-report items but some are of the performance type. It takes about 50 minutes to respond to all items. It is possible to respond to only some of them at one time and then return to the place in the test where responding was temporarily interrupted. It is also possible to use just a subset of the scales in order to shorten the test.

There is a norm group consisting of about 2000 persons. It can be broken down into subgroups based on gender, age and education. Separate norms are available for candidates for management positions and leader development. All scales are reliably measured (mean reliability = 0.75) and have been construct validated (mean construct validity = 0.68). In order to validate the test, external criteria have also been investigated, with success. It is hard to fake test results on *UPP* because impression management (IM) scales are used to correct for attempts to exaggerate the results. All these properties are discussed in detail in subsequent sections.

UPP measures a number of personality traits of importance in applied work:

- Social ability
- Emotional intelligence
- Willingness to cooperate
- Persistence
- Positive attitude
- Creativity
- Perfectionism
- Self esteem
- Narcissism
- Social confidence
- Independence

It also measures the Big Five personality dimensions:

- Emotional stability
- Extraversion
- Conscientiousness
- Openness
- Agreeableness

These 16 (11+5) basic scales are used to construct aggregate measures of:

[1] *UPP* stands for Understanding Human Potential.

- Ego strength
- Stress resilience
- Managerial ability
- Sales potential

Aggregate scales can also be constructed for special applications to measure both the presence of desirable traits and the absence of undesirable traits. In addition, *UPP* measures adjustment to the current work situation of the test taker. The factors measured are:

- Willingness to work
- Work interest
- Job satisfaction
- Willingness to work with changes
- Results orientation
- Work-life balance
- Control orientation
- Economic motivation

See Appendix for sample items of all scales.

Reliability

Cronbach's alpha values, mean intercorrelations among items of each scale, and variance accounted for by the two IM scales are given in Table 1.

Table 1. Basic psychometric properties of the *UPP* test scales.

Scale	Number of items or variables (aggregated measures)	Reliability	Mean intercorrelations of items in the scale	Proportion of variance accounted for by IM (two IM scales)	Proportion of variance accounted for by the FFM
Extraversion	10	0.86	0.49	0.139	-
Agreeableness	11	0.67	0.21	0.258	-
Emotional stability	8	0.75	0.34	0.276	-
Openness	9	0.67	0.19	0.104	-
Conscientiousness	14	0.75	0.20	0.098	-
Persistence	8	0.78	0.23	0.112	0.492
Willingness to cooperate	11	0.79	0.18	0.243	0.459
Positive attitude	10	0.84	0.28	0.338	0.364
Self esteem	9	0.75	0.29	0.273	0.522
Social ability	8	0.78	0.23	0.153	0.481
Emotional intelligence, self report items	15	0.75	0.15	0.207	0.415
Emotional intelli-	32	0.75	0.45	0.000	0.021

Table 1. Basic psychometric properties of the *UPP* test scales.

Scale	Number of items or variables (aggregated measures)	Reliability	Mean intercorrelations of items in the scale	Proportion of variance accounted for by IM (two IM scales)	Proportion of variance accounted for by the FFM
gence, performance, emotion identification judgments					
Creativity	8	0.75	0.33	0.023	0.332
Perfectionism	10	0.77	0.26	0.042	0.352
Social confidence	8	0.70	0.23	0.203	0.018
Narcissism	8	0.79	0.33	0.049	0.034
Independence	9	0.64	0.17	0.012	0.087
Job satisfaction	3	0.92	0.77	0.101	0.114
Willingness to work	7	0.85	0.44	0.136	0.167
Results orientation	13	0.70	0.11	0.058	0.330
Willingness to work with changes	8	0.70	0.09	0.071	0.295
Work interest	6	0.78	0.42	0.106	0.125
Control orientation	8	0.86	0.42	0.045	0.197
Work-life balance	9	0.90	0.53	0.098	0.135
Economic motivation	8	0.69	0.22	0.124	0.041
Impression management 1 (covert)	12	0.69	0.11	-	0.573
Impression management 2 (overt)	9	0.71	0.25	-	0.363
- Aggregate variables -					
Ego strength	10	0.68	0.16	-	
Managerial ability	13	0.80	0.24	-	
Stress resilience	7	0.61	0.18	-	
Sales potential	12	0.66	0.15	-	
Mean		0.758		0.131	0.269

The aggregate variables were corrected for IM. The median reliability of the scales, excluding the aggregated variables, is 0.76. Over a period of six weeks, test scale scores correlated on the average 0.70 between the two occasions. The median proportion of variance accounted for by IM is 0.13, corresponding to a correlation of 0.36, with a large variation among test scales, from 0 to 0.338. Since the Big Five accounted, on average, of 27 % of the variance, it can be concluded that error variance + Big Five variance is 24 % + 27 %, true variance not accounted for by measurement error or Big Five amounts to 49 %.

Construct validity

All test scales have been construct validated, in most cases by correlating them with internationally well-established and similar variables, which should be expected to be correlated with the *UPP* scales for other reasons. The correlations have been corrected for measurement error both in the criteria and the test scales. The results are given in Tables 2 and 3.

Table 2. Construct validation of basic and aggregated personality scales.

Variable	Construct criteria	Construct validity[2] (corrected for measurement errors)
Extraversion	Extraversion IPIP[3]	0.90
Emotional stability	Emotional stability IPIP	0.86
Openness	Openness IPIP	0.72
Agreeableness	Agreeableness IPIP	0.78
Conscientiousness	Conscientiousness IPIP	0.93
Willingness to cooperate	Hogan's passive aggression scale [36]	0.66
Creativity	Openness IPIP	0.68
Emotional intelligence	Schutte's et al. EI scale [70], Furnhams TEIQue [60]	0.59
Social ability	The UCLA loneliness scale [68], the Jones et al. shyness scale [38], UPP scale of social confidences	0.71
Positive attitude	Affect scales by Tsai et al. [95]	0.47
Persistence	Grit [19]	0.51
Self esteem	Rosenberg [65]	0.68
Social confidence	Jones et al. shyness scale [38]	0.96
Narcissism	NPI-16 [2]	0.80
Independence	Independence IPIP	0.63
Perfektionism	The CMD scale [26]	0.36
Ego strength	Hardiness [46] and proactivity [3]	0.69
Managerial potential	Leadership IPIP [83]	0.59
Stress resilience	Vulnerability IPIP [83]	0.31

Most of these values are quite satisfactory. The median construct validity is 0.68.

[2] In some cases these correlations are means of results from several studies

[3] http://ipip.ori.org/

Table 3. Construct validation of attitude (proxy criteria) scales.		
Variable	Construct criteria	Construct validity (corrected for measurement errors)
Willingness to work with changes	Oreg's scale of resistance to change [56; 57]	0.82
Willingness to work	ALI job satisfaction scale [33]	0.73
Results orientation	Ray's scale of achievement motivation [62; 63]	0.75
Work interest	The BPS (boredom) scale [100]	0.55
Job satisfaction	A pooled measure of the three work motivation dimensions by Deci and Ryan [17]	0.71
Control orientation	Burger's scale for measuring desire for control [7]	0.60
Ekonomic motivation	Self estimate of financial success [83]	0.55
Work-life balance	Hayman's balance scale [32]	0.76

The median construct validity of these variables is 0.72.

Proxy validation

Some of the test dimensions can be used as proxy criteria in validating the test. Dimensions such as willingness to work are credible proxies for job performance. Proxy criteria greatly simplify and encourage test validation. Table 4 shows representative values for correlations between some proxy criteria and job performance.

Table 4. Correlations between proxy criteria in the *UPP* test and job performance in a study of police officers [16].	
Proxy criterion	Correlation with job performance
Job satisfaction	0.37
Willingness to work	0.44
Job interest	0.37

Several studies have shown that proxy validity tends to be closely related to traditional external validity. Fig. 2 gives an example from a study of police officers [16].

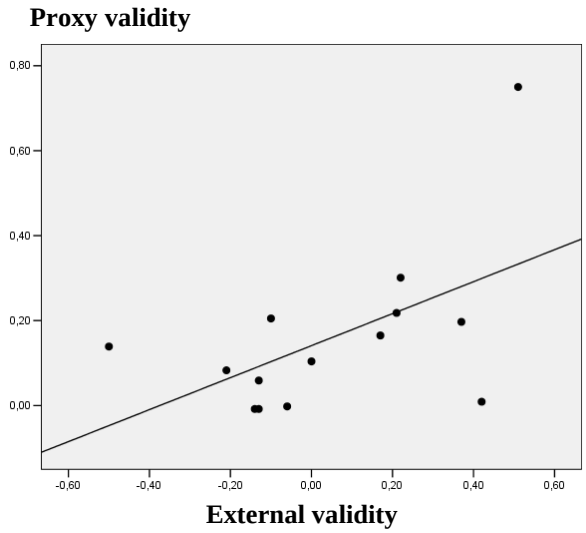

Figure 2. Proxy validity of UPP test scales plotted against external validity. The number of interrogations carried out during one year as external criterion.

Proxy validation greatly simplifies comparative studies of different types of prognostic information, e.g. different personality tests. In one study, *UPP* was compared to CTI [8] in the selection of candidates for officer training in the Swedish armed forces [88]. CTI is a self-report personality test which is very different from *UPP*. It is based on Jung's theory of personality types [40] and also includes psychopathy scales. It was found that *UPP* had a much stronger relationship to a proxy criterion than did CTI.

A further validation of *UPP* is reported here, using proxy criteria. See Table 5.

Table 5. Validation against 6 proxy criteria (squared multiple correlations), norm data.

Criterion	The FFM model	The *UPP* test
Willingness to work with changes	0.137	0.274
Job satisfaction	0.008	0.454
Willingness to work	0.010	0.475
Result orientation	0.106	0.212
Work interest	0.054	0.389
Balance life/work	0.045	0.097
Mean	0.041	0.317

The improvement in validity beyond the FFM was highly statistically significant in all cases, and dramatically large. A mean explained variance of 0.317 corresponds to a criterion correlation of 0.56, very close to other data presented later in this paper and much better than the common results for tests of the first and second generation personality tests, which tend to lie in the interval 0.2 – 0.3.

Emotional intelligence

EI is a somewhat controversial concept, and many personality tests do not cover it. The concept of EI has, however, many interesting implications [20; 22; 74; 89], among them a negative correlation with Machiavellianism, suggesting that EI can be used as a proxy for that important part of "the dark triad" [55]. EI has been found to provide incremental validity above what is achieved by the FFM [39].

Research on EI is extensive and increasing, see Fig. 3. The present section gives some details about EI measurement in *UPP*, and a few selected results.

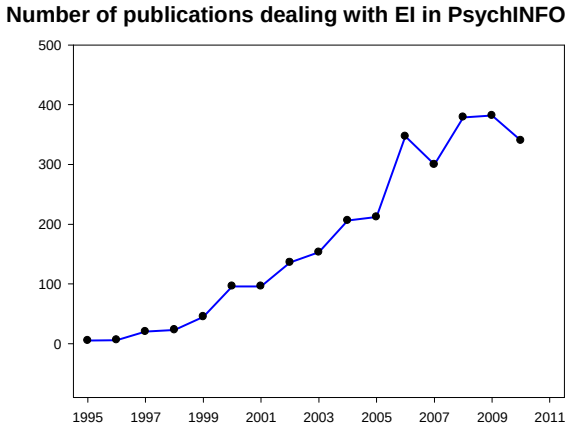

Figure 3. Development of number of scientific publications about EI.

Most of the scales used in the *UPP* test use a self-report format, including an EI scale. However, a special section includes the task of identifying emotions in facial expressions of the type depicted in Fig. 4.

Fig. 4. Types of facial expressions used in the *UPP* test.

The facial expressions are judged on eight emotion scales; consensus is used to define the "correct" answer [74; 77]. The relation between self-report and performance scales of emotional intelligence in *UPP* is fairly weak, much like the situation in research on these topics [39]. In a group of 725 candidates for manager positions, the correlation between the self-report scale and the performance scale was 0.11 (p = 0.003). However, both types of scales carry important information [21], and they tend to agree across groups if not at the level of individuals. See Fig. 5 for the relationship between EI and age.

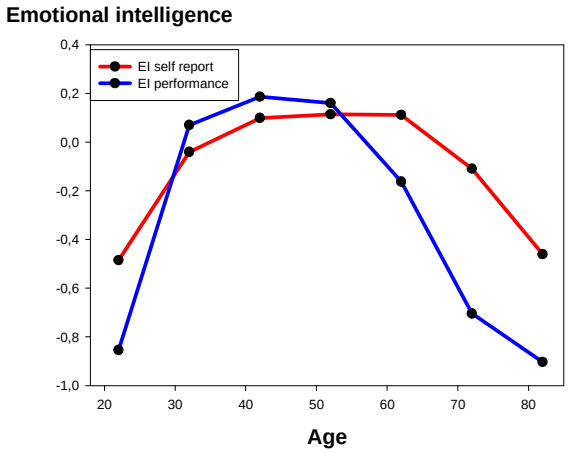

Figure 5. Relation between age and EI for self-report and performance scales, norm data.

A gender difference in emotional intelligence is illustrated in Fig. 6. It is well known that women tend to have a higher emotional intelligence than men [39]. Fig. 7 shows differences in emotional intelligence as a function of civil status.

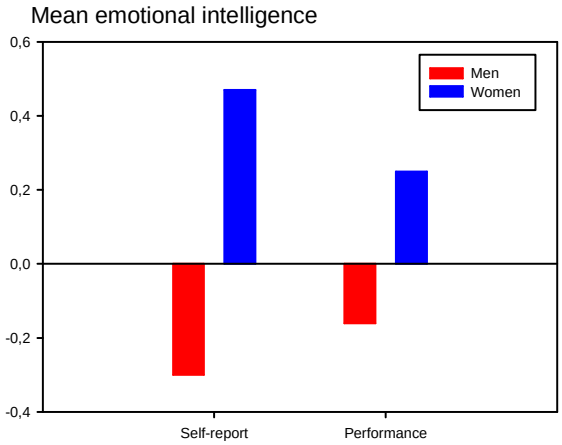

Figure 6. Gender and emotional intelligence

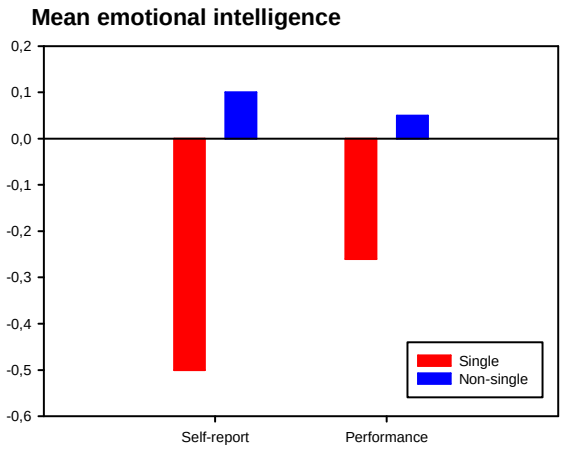

Figure 7. Civil status and emotional intelligence.

The figures illustrate that performance and self-report scales give similar results across age and gender groups, in the latter case also in the expected direction.

Correcting for IM in the UPP test

A social desirability scale, or lie scale, can be used in selection situations for weeding out those that have high value, but that strategy does not imply that the effect of impression management disappears, it is only somewhat mitigated [76]. The lie scale is typically used globally for an entire test, while research has demonstrated that impression management is of very different importance depending on the test variable being studied. Table 6 gives an example from a study of a large data set[4] from job seekers who took the *UPP* test (the screening module) in connection with applying for a job [78]. Correlations between personality variables and a measure of impression management are apparently quite variable. Hence, a global approach to correcting for IM will be insufficiently effective. It will over-correct for some scales and under-correct for others.

Table 6. Proportion of variance accounted for by social desirability responding, N=2202.

Test variable	Proportion of variance explained by tactical responding
Extraversion	0.213
Endurance	0.398
Will to cooperate	0.419
Positive basic attitude	0.381
Creativity	0.089

[4] These and subsequent results summarized in the present paper come from studies documented in (Sjöberg, 2008a) and the *UPP* test manual, available for download at http://www.psykologisk-metod.se/files/manual%20komplett%20februari%202010.pdf.

In the *UPP* test, a regression model is fitted to each test variable separately and residuals are used to estimate corrected scale values. Fig. 8 shows the results for a sample of job applicants and the norm group. Before correction, there was a large difference between job applicants and norm data. After correcting for IM the difference was dramatically reduced. In several studies, we have found that about 90 percent of the effect of IM is eliminated in this manner. The approach has been validated both experimentally and on real job application data.

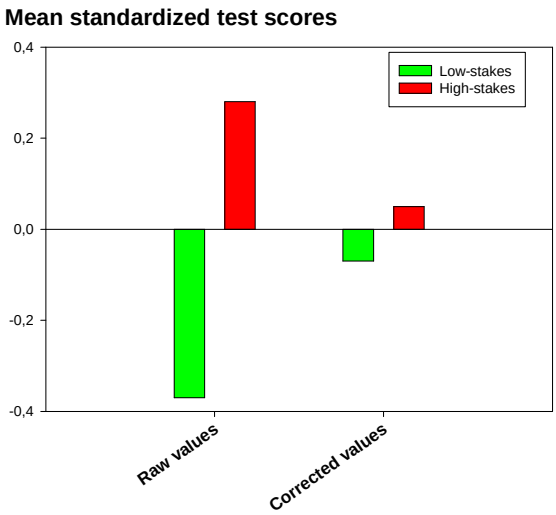

Figure 8. Mean values of personality dimensions before and after correction for IM, high-stakes (job applicants) and norm data.

Fig. 9 gives corresponding results from an experimental study where some participants were instructed to fake good, others just to answer in an honest manner.

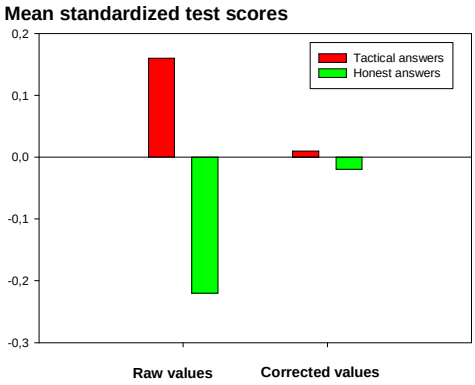

Figure 9. Mean values of personality dimensions before and after correction for IM, experimental study.

It is interesting to note that gender differences in applications for management jobs, favoring men over women, are greatly reduced due to IM correction, see Fig. 10.

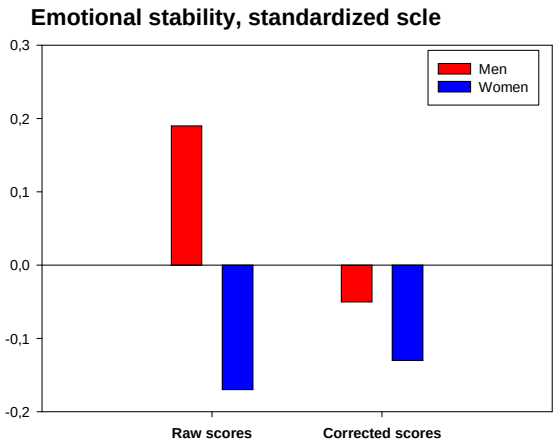

Figure 10. Raw and corrected scores in emotional stability of men and women applying for management jobs.

Clearly, if women answer in a more honest way than men do they will be at a disadvantage in career contexts. The *UPP* test counteracts the drawbacks of female honesty. No other test does so, as far as is known. In one study, it was also found that immigrants were at a similar disadvantage when having taken a personality test, a disadvantage which was eliminated by our method for correcting for IM [82].

In our current research, we compared applicants and incumbents in the officer's training program in the Swedish Armed Forces. Fig. 11 shows IM scale means for the two groups.

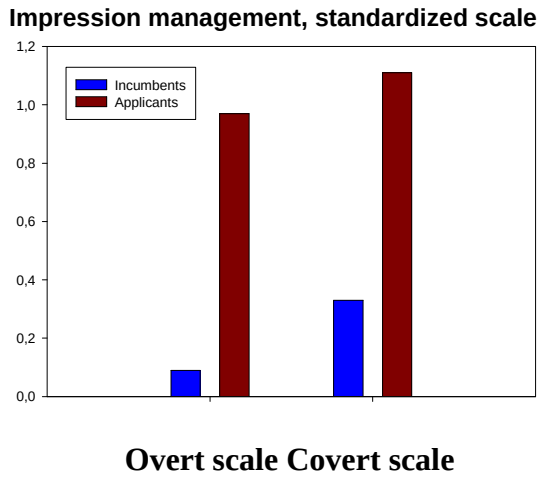

Figure 11. IM in applicants and incumbents, the officer's training program. Norm data set to mean=0

The figure shows *very large* [12] average differences between applicants and incumbents. As expected, applicants were more tactical in their test responses. Fig. 12 shows differences in emotional stability before and after correcting for IM. Similar results were obtained for other test scales.

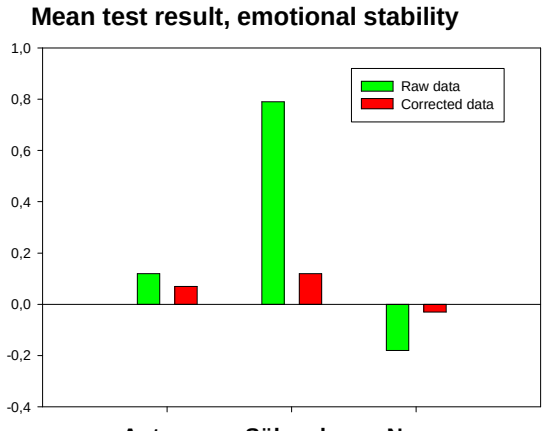

Figure 12. Mean emotional stability for applicants and incumbents, before and after correcting for IM.

How is validity affected by the method used for correcting for IM? Validities of the test scales were increased after correction for IM, se Fig. 13.

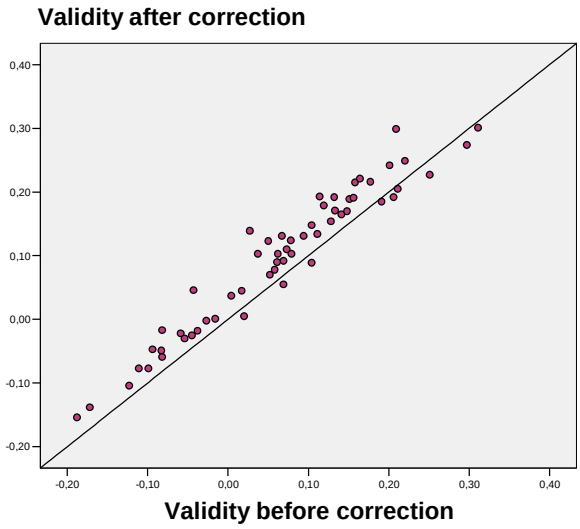

Figure 13. Validities of UPP test scales before and after correction for IM.

These data were from a study of middle managers in medical care whose leadership was assessed in 360-degrees judgments by their co-workers [86]. Each point in the graph corresponds to a test scale and a criterion measure. It is seen that validities after correction for IM were higher than those obtained before correction.

In conclusion, the methodology applied in the *UPP* test worked well and removed most of the effects of tactical responding to the test. Initial, extremely large, differences between applicants and incumbents were eliminated.

Mood and test results

It is likely that responses on a personality test are affected by the current mood state of the test-taker, but this is seldom attended to in practical testing. Table 7 shows correlations between Big Five test dimensions and mood state at the time of testing. Mood was measured with a scale constructed by Sjöberg, Svensson and Persson [91], which has been widely used. The data are from testing applicants to the Stockholm School of Economics.

Table 7. Correlations between Big Five personality scales and current mood state, high-stakes testing, N=210.

	Mood dimension			
Personality scale	Happy-sad	Alert-tired	Calm-tense	Pooled measure of mood
Agreeableness	0.09	0.11	0.08	0.11
Emotional stability	0.50**	0.35**	0.38**	0.50**
Openness	0.38**	0.50**	0.21**	0.41**
Extraversion	0.43**	0.37**	0.22**	0.39**
Conscientiousness	0.01	0.30**	0.03	0.11

** p<0.01

The table shows that three of the five dimensions were rather strongly correlated with mood. Extreme groups show the results even more clearly, see Fig. 14, where the 10 percent worst mood cases are compared with the rest of the test-takers in standardized scores.

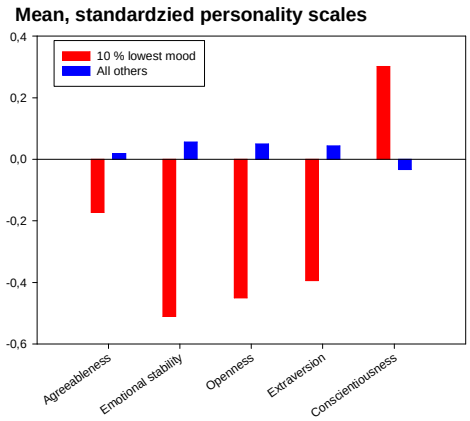

Figure 14. Mean personality scales values for the 10 % of the test takers who were in the worst mood, compared to all others. Data from a high-stakes situation.

The data suggest that a low mood can lead to distorted values on a personality scale. For this reason, the *UPP* test includes scales for measuring mood. This is an aspect of data quality, and a low mood constitutes a warning signal. Possibly, it should lead to repeated testing at a

later occasion, or to caution in interpreting the test results. Other aspects of data quality are treated in the following section.

Data quality

Data quality is important for using the results of a personality test. Some people are not careful when responding, others do not understand the task and the items as intended. Impression management is ever present.

UPP uses the following quality indices:

- Acquiescence (negative indicator)
- Intra-individual variability (positive indicator)
- Similarity of responses in relation to group means (positive)
- Social desirability scale, overt (negative)
- Social desirability scale, covert (negative)

These indices were correlated, suggesting that they could be used to construct a pooled measure of data quality. It was found that data quality was:

- Higher for women than for men
- Higher for older than younger people
- Higher for test takers with a higher level of education
- Higher under high-stakes testing than low-stakes testing

Fig. 15 shows the relationship between mean data quality and level of education, separately for men and women.

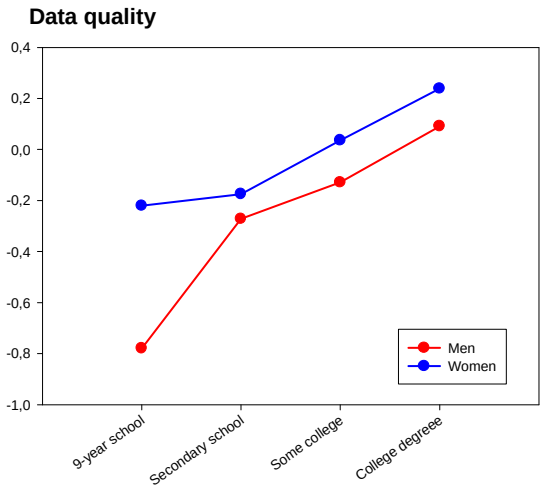

Figure 15. Mean data quality for men and women at different levels of education.

Errors in predicting a pooled measure of the proxy criteria (absolute scores) from the personality scales (multiple regression) are plotted against the index of data quality in Fig .16.

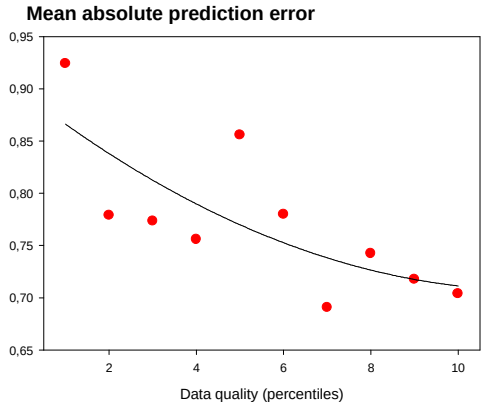

Figure 16. Mean absolute prediction error plotted against data quality (percentile groups).

A low value of data quality is a warning signal, and the test should be taken anew, or disregarded.

Attitude toward the test: “Face validity”

It is important to know about the test takers’ attitudes toward the test, which is sometimes called face validity. A negative evaluation of the test by the test taker is an indication that something went wrong and should be followed up in an interview and possibly a renewed testing. The *UPP* test, therefore, is concluded with eight questions measuring attitude toward the test; these questions are correlated and are used to estimate a pooled measure of attitude. The distribution of attitude ratings for 770 test takers is provided in Fig. 17.

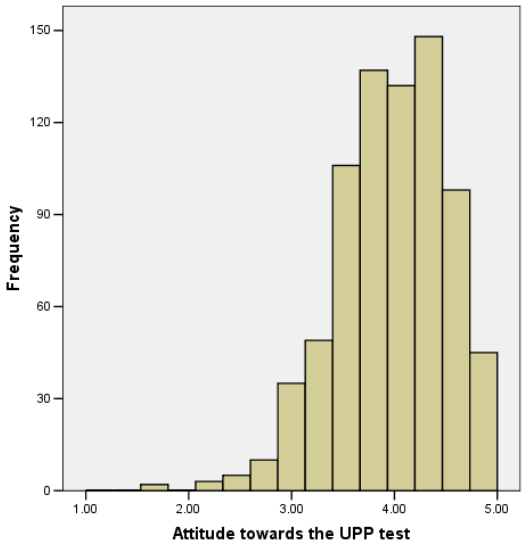

Figure 17. Distribution of mean ratings of attitudes toward the *UPP* test.

The figure shows an overwhelmingly positive attitude toward the test. It is usually not known how test taker have reacted to other tests. An exception. People dislike the ipsative formats which are so often used [31], e.g. by OPQ and PPA.

Validation of UPP test against external criteria

Principles

All subscales of a test cannot be expected to be relevant in all circumstances. The validity of a test is specific to the situation and the criteria. It is therefore misleading to assess the value of a test on the basis of the average, or median, correlation between all test scales and criteria, as done in the EFPA test assessment system[5] (also used in Sweden). A test could have a high level of validity if a few of its subscales correlate with a criterion even of several others do not.

The scales most likely to be predictive of a certain criterion are the ones showing the highest degree of similarity with it [34]. This is an example of a very general principle for behavioral data: correlations and similarity tend to converge [72].

Selection of scales could also be based on multiple regression but sample sizes are rarely sufficient. Unit weights applied to a few selected scales will usually function quite well [5].

Social job skills

Social job skills was measured in one study [90] and related to *UPP*, see Table 8.

Table 8. Hierarchic regression analysis of social job skills related to the *UPP* tests.

Block of independent variables	R^2_{adj}	ΔR^2	F for ΔR^2	df	*p*
Step 1: The FFM model	0.105	-	-		0.006 (The FFM model)
Step 2: FFM + emotional intelligence and social ability	0.323	0.221	17.320	2,99	<0.0005 (Added explanatory power)

The level of explained variance reached, 32.3 %, corresponds to a correlation of 0.57 with the criterion. The two pertinent *UPP* variables contributed dramatically better than the FFM model.

Management career

Management career was related to the *UPP* test [79], see Table 9. The criterion was binary: promoted to manager or not.

[5] http://www.efpa.eu/

Table 9. Hierarchical binary regression analysis of manager responsibility against the *UPP* test. N=107.

Block	*R*, Cox & Snell	*R*, Nagelkerke	χ^2	df	*p*
The entire *UPP* test	0.412	0.516	18.968	3	<0.0005
The FFM model variables	0.045	0.055	0.172	1	0.678
UPP variables beyond FFM	0.410	0.507	18.280	2	<0.0005

The criterion validity of the *UPP* is very satisfactory in these data. The most realistic estimate of the correlation is 0.52. Three more studies have reported results on external validation, see next section.

Managers in medical care

A study of 166 middle managers in medical care in the Stockholm county [86]. In this study, indices were formed to be similar to three dimensions of 360-degrees judgments, focusing on supporting subordinates, promoting production or change. Correlation between CPE data (360-degrees) and matched UPP indices are given in Table 10.

Table 10. Correlations between matched *UPP* indexes and CPE dimensions, after correction for measurement error and restricted range in test variables

Evaluator	Employees	Production	Change
Superior manager	0.50**	0.15	0.39*
Peers	0.19	0.12	0.30*
Subordinates	0.39*	0.33*	0.46***
Self-assessment	0.53***	0.62***	0.72***

*** p<0.001 ** p<0.01 * p<0.05

The average validity for judgments made by others was around 0.5, and higher for self-assessments.

Service quality in a finance company

In this study, 53 persons employed in customer service (of a total of 65) in a finance company took the *UPP* test [81]. Their work performance was rated by their supervisors and related to the test results. The most valid test scales were:

- positive attitude
- persistence
- emotional stability
- extraversion
- willingness to cooperate
- work-life balance

A pooled index with equal weights to these six variables resulted in the following correlations with criteria:

- 0.50 with the assessed over-all value to the firm
- 0.38 with performance in core job tasks
- 0.28 with social support given to work mates

It is interesting to note that the highest correlation was that with perceived over-all value to the firm.

Police authorities

This is a study of 100 employees of police authorities [16]. They took *UPP* and had been assessed by superiors for salary decisions. The latter judgments were pooled into two overarching dimensions: productivity and social function at work. The correlation with an index consisting of the scales selected in the study of service functions was 0.50 for productivity and 0.38 for social functions. Productivity measured as the number of completed interrogations in one year correlated 0.54 with the test index.

A study of applicants to a Police Academy showed that assessments of the applicants correlated with the *UPP* test at the 0.3 level [85]. Such judgments are typically difficult to predict with self-report personality tests.

Dealing with test complexity

People have a limited ability to make complex judgments without the support of computers and explicit decision rules. This fact has been well-known for many years. An often cited classic is a paper by Miller [51]. Expert judgments of many kinds, including the assessment of job applicants, have confirmed this general principle [15; 30]. There are some interesting exceptions in special cases, if the experts get fast and clear feedback based on valid theory [42]. These conditions are rarely present in the assessment of job applicants.

It is usual for judges to come to different conclusions if the information they use is complex and extensive - a common situation. Furthermore, assessments tend to vary over time. At the same time that we have these limitations in our judgment capacity, we have a tendency to fall prey to an illusion. The more information we get, the more confident we are - but beyond a modest limit, judgments become worse as in formation increases. See Fig. x.

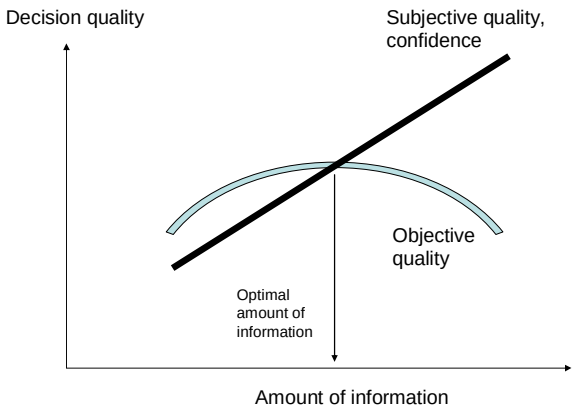

Figure 18. Decision quality as a function of amount of information.

Most personality tests give a complicated picture of a person. This is reasonable since everyone "knows" that people are complicated. Popular tests provide results for 30-40 dimensions. It is likely that such abundance of information is popular due to the information illusion discussed above. More information makes us more confident. Research has, however, shown that explicit rules for combining formation gives better results. Such a rule can simply be based on the decision maker's own systematic strategy, so-called boot-strapping [29], or explicitly judged importance weights. The use of weights is an effective way of answering the question: "How do I interpret this test result?" The alterative approach is use a holistic evaluation based on the pattern of results. Holism has traditionally had a strong position in the interpretation of test results, but it cannot be justified on empirical and scientific grounds [67].

Subjective interpretation typically results in narrative texts which may be very credible, due to a number of psychological factors. Such factors have been discussed as enabling "cold reading", i.e. credible inferences about a person, which lack factual basis [66]. Historical examples show how credibility of the Rorschach test was established by "wizards" who could seemingly produce surprisingly correct statements about a person on the basis of responses to that test [101], in spite of the fact that this test, as well as other projective techniques have been found to lack validity [27; 43]. I give two examples of research, which illustrate how illusory credibility may be established.

The Forer effect. Flattering texts, which are full of statements which are generally true and which say "both A and its Opposite B" are perceived as very accurate. Forer showed this in a classic study a long time ago [25] ; results which have been replicated many times [18; 94].

Forer gave a group of students a "test" which he said would reveal their personalities. After some time a returned with narrative texts said to be based on the responses to the test. Each students got his or her text, but they were all the same. They were asked to judge how well the texts described their personalities. About 90 % said that the texts fitted very well. Here is what they got (typical astronomical texts):

> *"You have a need for other people to like and admire you, and yet you tend to be critical of yourself. While you have some personality weaknesses you are generally able to compensate for them. You have considerable unused capacity that you have not turned to your advantage. Disciplined and self-controlled on the outside, you tend to be worrisome and insecure on the inside. At times you have serious doubts as to whether you have made the*

right decision or done the right thing. You prefer a certain amount of change and variety and become dissatisfied when hemmed in by restrictions and limitations. You also pride yourself as an independent thinker; and do not accept others' statements without satisfactory proof. But you have found it unwise to be too frank in revealing yourself to others. At times you are extroverted, affable, and sociable, while at other times you are introverted, wary, and reserved. Some of your aspirations tend to be rather unrealistic. "

MBTI and PPA excel in using statements of this type , and they provide popular reading for those who have taken the tests. They are perceived to be almost perfectly accurate and to give self insights, but they simply flatter [93] and/or confirm already existing self beliefs. Once credibility is established the tester can give important advice about selection, team composition and personal development. No research exists, which shows such advice to be useful, but since the test report is so persuasive the advice is probably also believed.

The *"Draw-a-man"-effect".* The draw-a-man test is credible to many users although it has no demonstrated validity [98]. This is because of common-sense thinking about what various aspect of a drawing could mean. Example: large muscles mean problem with male self-image, large eyes imply paranoid tendencies, etc. Inn addition, there is selective memory of cases which supported these speculations, the others are forgotten or explained away [10; 11].

The UPP test deals with complexity with aggregate variables, which are linear composites of selected subscales. Extensive research, over a period of 50 years, has shown that this approach is superior to subjective integration of information [30; 50].

Conclusions

UPP has a number of unique advantages:

- The effect of impression management (IM) are very salient but can be eliminated by about 90 percent. Correction for IM is done for each scale separately and is empirically based.

- The 11 focused scales (see above) make possible a dramatically improved validity when compared to the traditional Big Five, about eight times better.

- The scales measuring adjustment to the current work situation can also be used as proxy criteria for the evaluation of *UPP* or other tests, and in co-worker studies, giving a unique chance to get a psychologically more informative view than in usual surveys.

- The test assesses several aggregate variables, which summarize the results on subscales in order to simplify and improve inferences and interpretations.

- The test also assesses the quality of test data and mood, and gives a warning signal when quality is low and a re-testing is called for.

- Data are collected on the test takers' evaluation of the test, giving a measure of "face validity".

UPP is flexible in the sense that scales can be deleted if a shorter test desired, and some scales are not relevant in a given application. *UPP* is also available in two short forms intended for screening purposes. Impression management is always applied.

References

[1]. Allport, F. H., & Allport, G. W. (1921). Personality traits: Their classification and measurement. *Journal of Abnormal and Social Psychology, 16*, 6-40.

[2]. Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. [doi:10.1016/j.jrp.2005.03.002]. *Journal of Research in Personality, 40*, 440-450.

[3]. Bateman, T. S., & Crant, J. M. (1993). The proactive component of organizational behavior: A measure and correlates. *Journal of Organizational Behavior, 14*, 103-118.

[4]. Bergner, S., Neubauer, A. C., & Kreuzthaler, A. (2010). Broad and narrow personality traits for predicting managerial success. [doi:10.1080/13594320902819728]. *European Journal of Work and Organizational Psychology, 19*, 177-199.

[5]. Bobko, P., Roth, P. L., & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores: A literature review, application to content validity, and meta-analysis. [doi:10.1177/1094428106294734]. *Organizational Research Methods, 10*, 689-709.

[6]. Brown, A., & Bartram, D. (2009). *Development and psychometric propertioes of the OPQ32r. Supplement to the OPQ 32 technical manual*: SHL.

[7]. Burger, J. M. (1992). *Desire for control: Personality, social, and clinical perspectives*: New York, NY, US: Plenum Press.

[8]. Carlstedt, L., & Widén, H. (1998). *CTI – Commander Trait Inventory. Manual*. Karlstad: LI: Försvarshögskolan.

[9]. Cattell, R. B. (1948). Primary personality factors in the realm of objective tests. *Journal of Personality, 16*, 459-487.

[10]. Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abormal Psychology, 73*, 193-204.

[11]. Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*, 271-280.

[12]. Cohen, J. (1988). *Statistical power analysis for behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

[13]. Cooper, C. (2002). *Individual differences. 2nd edition*. London: Hodder Education.

[14]. Costa, P. T., Jr, & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.

[15]. Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668-1674.

[16]. de Colli, D. (2011). *Ett nytt svenskt arbetspsykologiskt test och arbetsprestation inom polisen – samtidig validitet*: Mälardalens högskola, Akademin för hållbar samhälls- och teknikutveckling.

[17]. Deci, E. L., & Ryan, R. M. (1991). A motivational approach to the self: Integration in personality. In R. Dienstbier (Ed.), *Nebraska symposium on motivation: Vol. 38. Perspectives on motivation* (pp. 237-288). Lincoln: University of Nebraska Press.

[18]. Dickson, D. H., & Kelly, I. W. (1985). The 'Barnum Effect in Personality Assessment: A Review of the Literature. *Psychological Reports 57*, 367-382.

[19]. Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*, 1087-1101.

[20]. Engelberg, E., & Sjöberg, L. (2004). Internet use, social skills and adjustment. *CyberPsychology & Behavior, 7*, 41-48.

[21]. Engelberg, E., & Sjöberg, L. (2005). Emotional intelligence and interpersonal skills. In R. D. Roberts & R. Schulze (Eds.), *International handbook of emotional intelligence* (pp. 289-308). Cambridge MA: Hogrefe.

[22]. Engelberg, E., & Sjöberg, L. (2006). Money attitudes and emotional intelligence. *Journal of Applied Social Psychology, 36*, 2027-2047.

[23]. Eysenck, H. J. (1959). *Manual of the Maudsley Personality Inventory*. San Diego, CA: Educational and Industrial Testing Service.

[24]. Ferguson, E., Payne, T., & Anderson, N. (1994). Occupational personality assessment: Theory, structure and psychometrics of the OPQ FMX5-Student. *Personality & Individual Differences, 17*, 217-225.

[25]. Forer, B. R. (1949). The fallacy of personal validation: a classroom demonstration of gullibility. *Journal of Abnormal & Social Psychology, 44*, 118-123.

[26]. Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research, 14*, 449-468.

[27]. Garb, H. N., Lilienfeld, S. O., & Wood, J. M. (2004). Projective techniques and behavioral assessment. In S. N. Haynes & E. M. Heiby (Eds.), *Comprehensive handbook of psychological assessment, Vol. 3: Behavioral assessment* (pp. 453-469). Hoboken, NJ, US: John Wiley & Sons Inc.

[28]. Gilliland, S. W. (1995). Fairness from the applicants perspective - reactions to employee selection procedures. *International Journal of Selection and Assessment, 3*, 11-19.

[29]. Goldberg, L. R. (1970). Man versus model of man: A rationale plus some evidence for a method of improving clinical inferences. *Psychological Bulletin, 73*, 422-432.

[30]. Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293-323.

[31]. Harland, L. K. (2003). Using Personality Tests in Leadership Development: Test Format Effects and the Mitigating Impact of Explanations and Feedback. *Human Resource Development Quarterly, 14*, 285-301.

[32]. Hayman, J. (2005). Psychometric assessment of an instrument designed to measure work life balance. *Research and Practice in Human Resource Management, 13*, 85-91.

[33]. Hellgren, J., Sjöberg, A., & Sverke, M. (1997). Intention to quit: Effects of job satisfaction and job perceptions. In F. Avallone, J. Arnold & K. d. Witte (Eds.), *Feelings work in Europe* (pp. 415-423). Milano: Guerini.

[34]. Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: A socioanalytic perspective. *Journal of Applied Psychology, 88*, 100-112.

[35]. Hogan, R. (1992). Hogan Personality Inventory. *Psychological Test Bulletin, 5*, 130-136.

[36]. Hogan, R., & Hogan, J. (1997). *Hogan Development Survey.*: Hogan Assessment Systems.

[37]. Irvine, S. H., & Lindelöw-Danielsson, M. (2004). *PPA validity studies in Scandinavia: An occupational validity study in Sweden*. Stockholm: SLG International.

[38]. Jones, W. H., Briggs, S. R., & Smith, T. G. (1986). Shyness: Conceptualization and measurement. [doi:10.1037/0022-3514.51.3.629]. *Journal of Personality and Social Psychology, 51*, 629-639.

[39]. Joseph, D. L., & Newman, D. A. (2010). Emotional intelligence: An integrative meta-analysis and cascading model. [doi:10.1037/a0017286]. *Journal of Applied Psychology, 95*, 54-78.

[40]. Jung, C. G. (1921/1971). *Psychological types*. Princeton NJ: Princeton University Press.

[41]. Jung, C. G. (1950). *Psychologische Typen, 8. aufl.* Zürich: Rascher.

[42]. Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. [doi:10.1037/a0016755]. *American Psychologist, 64*, 515-526.

[43]. Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27-66.

[44]. Lounsbury, J. W., Sundstrom, E., Loveland, J. L., & Gibson, L. W. (2002). Broad versus narrow personality traits in predicting academic performance of adolescents. [doi:10.1016/j.lindif.2003.08.001]. *Learning and Individual Differences, 14*, 65-75.

[45]. Mabon, H. (2004). *Arbetspsykologisk testning. Om urvalsmetoder i arbetslivet. Andra upplagan*. Stockholm: Psykologiförlaget.

[46]. Maddi, S. R. (2006). Hardiness: The courage to grow from stresses. *The Journal of Positive Psychology, 1*, 160-168.

[47]. Marston, W. M. (1989/1928). *Emotions of normal people*. Ormskirk, Lancs.: Thomas Lyster.

[48]. Matthews, G., Emo, A. K., Funke, G., Zeidner, M., Roberts, R. D., Costa Jr, P. T., et al. (2006). Emotional intelligence, personality, and task-induced stress. [doi:10.1037/1076-898X.12.2.96]. *Journal of Experimental Psychology: Applied, 12*, 96-107.

[49]. McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

[50]. Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota Press.

[51]. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

[52]. Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

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

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

[55]. O'Boyle Jr, E. H., Forsyth, D. R., Banks, G. C., & McDaniel, M. A. (2011). A meta-analysis of the dark triad and work behavior: A social exchange perspective. [doi:10.1037/a0025679]. *Journal of Applied Psychology*No Pagination Specified.

[56]. Oreg, S. (2003). Resistance to change: Developing an individual differences measure. *Journal of Applied Psychology, 88*, 680-693.

[57]. Oreg, S., Bayazit, M., Vakola, M., Arciniega, L., Armenakis, A., Barkauskiene, R., et al. (2008). Dispositional resistance to change: Measurement equivalence and the link to personal values across 17 nations. *Journal of Applied Psychology, 93*, 935-944.

[58]. Paul, A. M. (2004). *The cult of personality. How personality tests are leading us to miseducate our children, mismanage our companies, and misunderstanding ourselves*. New York: Free Press.

[59]. Paunonen, S. V., Haddock, G., Forsterling, F., & Keinonen, M. (2003). Broad versus narrow personality measures and the prediction of behaviour across cultures. *European Journal of Personality, 17*, 413-433.

[60]. Pérez, J. C., Petrides, K. V., & Furnham, A. (2005). Measuring trait emotional intelligence. In R. Schulze & R. D. Roberts (Eds.), *International Handbook of Emotional Intelligence*. Cambridge, MA: Hogrefe & Huber.

[61]. Ramfalk, C. W. (1957). *A study of the selection of personnel for higher positions in industrial organizations*. Stockholm: Almqvist & Wiksell.

[62]. Ray, J. J. (1975). A behavior inventory to measure achievement motivation. *Journal of Social Psychology, 95*, 135-136.

[63]. Ray, J. J. (1979). A quick measure of achievement motivation: Validated in Australia and reliable in Britain and South Africa. *Australian Psychologist, 14*, 337-344.

[64]. Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational & Organizational Psychology, 66*, 225-244.

[65]. Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

[66]. Rowland, I. (2005). *The full facts book of cold reading, 4th edition*. London: Full Facts Books.

[67]. Ruscio, J. (2002). The emptiness of holism. *Skeptical Inquirer, 26*, 46-50.

[68]. Russell, D. (1996). The UCLA loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment, 66*, 20-40.

[69]. Sartori, R. (2010). Face validity in personality tests: Psychometric instruments and projective techniques in comparison. [doi:10.1007/s11135-009-9224-0]. *Quality & Quantity: International Journal of Methodology, 44*, 749-759.

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

[71]. SHL. (1999-2007). *OPQ32. SHL sales report. Technical manual*: SHL Group Ltd.

[72]. Sjöberg, L. (1980). Similarity and correlation. In E.-D. Lantermann & H. Feger (Eds.), *Similarity and choice* (pp. 70-87). Bern: Huber.

[73]. Sjöberg, L. (1989). Mood and expectation. In A. F. Bennett & K. M. McConkey (Eds.), *Cognition in individual and social contexts* (pp. 337-348). Amsterdam: Elsevier.

[74]. Sjöberg, L. (2001). Emotional intelligence: A psychometric analysis. *European Psychologist, 6*, 79-95.

[75]. Sjöberg, L. (2005). En kritisk diskussion av Myers-Briggs testet. (A critical discussion of the Myers-Briggs test). *Organisational Theory & Practice. Scandinavian Journal of Organisational Psychology, 15*, 21-28.

[76]. Sjöberg, L. (2008). *Bortom Big Five: Konstruktion och validering av ett personlighetstest. (Beyond Big Five: Construction and validation of a personality test)* (SSE/EFI Working Paper Series in Business Administration No. 2008:7). Stockholm: Stockholm School of Economics.

[77]. Sjöberg, L. (2008). Emotional intelligence and life adjustment. In J. C. Cassady & M. A. Eissa (Eds.), *Emotional Intelligence: Perspectives on Educational & Positive Psychology* (pp. 169-183). New York: Peter Lang Publishing.

[78]. Sjöberg, L. (2009). *UPP-testet: Korrektion för skönmålning. (The UPP test: Correction for impression management). Forskningsrapport 2009:3.* Stockholm: Psykologisk Metod AB.

[79]. Sjöberg, L. (2009). *UPP-testet: Kriterierelaterad validitet. Forskningsrapport 2009:2*. Stockholm: Psykologisk Metod AB.

[80]. Sjöberg, L. (2010). *Personlighetsdimensioners validitet i arbetslivet: teorier och empiri* (SSE/EFI Working Paper Series in Business Administration No. 2010:6). Stockholm: Stockholm School of Economics.

[81]. Sjöberg, L. (2010). *UPP-testet och kundservice: Kriteriestudie. Forskningsrapport 2010:6*. Stockholm: Psykologisk Metod AB.

[82]. Sjöberg, L. (2010). *UPP-testet: Mångfald gynnas av korrektion för skönmålning. Forskningsrapport 2010: 2*. Stockholm: Psykologisk Metod AB.

[83]. Sjöberg, L. (2011). *UPP och UPP/Screen i relation till motivation, personlighet och framgång*. Stockholm: Psykologisk Metod AB.

[84]. Sjöberg, L. (In press). Ipsativa och normativa svarsformat i personlighetstest. *Psykologtidningen*.

[85]. Sjöberg, L., & Annell, S. (2011). *UPP/Screen vid antagning till polisutbildningen. Rapport # 1*. Stockholm: Rekryteringsmyndigheten.

[86]. Sjöberg, L., Bergman, D., Lornudd, C., & Sandahl, C. (2011). *Sambandet mellan ett personlighetstest och 360-graders bedömningar av chefer i hälso- och sjukvården*. Stockholm: Karolinska Institutet, Institutionen för lärande, informatik, management och etik (LIME).

[87]. Sjöberg, L., & Biel, A. (1983). Mood and belief-value correlation. *Acta Psychologica, 53*, 253-270.

[88]. Sjöberg, L., Bäccman, C., & Gustavsson, B. (2011). *Personlighetstestning vid antagning till FHS officersutbildning. ILM Serie T:39, 2011*. Karlstad: Institutionen för ledarskap och Management, Försvarshögskolan.

[89]. Sjöberg, L., Littorin, P., & Engelberg, E. (2005). Personality and emotional intelligence as factors in sales performance. *Scandinavian Journal of Organizational Theory and Practice, 15*, 21-37.

[90]. Sjöberg, L., & Möller, K. (2010). *Sociala arbetsfunktioner och personlighet* (SSE/EFI Working Paper Series in Business Administration No. 2010:2). Stockholm: Stockholm School of Economics.

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

[92]. Tett, R. P., Steele, J. R., & Beauregard, R. S. (2003). Broad and narrow measures on both sides of the personality-job performance relationship. *Journal of Organizational Behavior, 24*, 335-356.

[93]. Thiriart, P. (1991). Acceptance of personality test results. *Skeptical Inquirer, 15*, 166-172.

[94]. Trankell, A. (1961). *Magi och förnuft i människobedömning*. Stockholm: Bonnier.

[95]. Tsai, J. L., Knutson, B., & Fung, H. H. (2006). Cultural variation in affect valuation. *Journal of Personality and Social Psychology, 90*, 288-307.

[96]. Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. [doi:10.1037/a0021196]. *Journal of Applied Psychology, 97*, 499-530.

[97]. Whetzel, D. L., McDaniel, M. A., Yost, A. P., & Kim, N. (2010). Linearity of Personality–Performance Relationships: A large-scale examination. *International Journal of Selection and Assessment, 18*, 310-320.

[98]. Willcock, E., Imuta, K., & Hayne, H. (2011). Children's human figure drawings do not measure intellectual ability. [doi:10.1016/j.jecp.2011.04.013]. *Journal of Experimental Child Psychology, 110*, 444-452.

[99]. Vitacco, M. J., Lilienfeld, S. O., Erickson, S., & Wood, J. M. (2012). Challenging personality testing: Objective and projective instruments *Coping with psychiatric and psychological testimony: Based on the original work by Jay Ziskin (6th ed)* (pp. 335-362). New York, NY, US: Oxford University Press.

[100]. Vodanovich, S. J. (2003). Psychometric measures of boredom: A review of the literature. *Journal of Psychology: Interdisciplinary and Applied, 137*, 569-595.

[101]. Wood, J. M., Nezworski, M. T., Lilienfeld, S. O., & Garb, H. N. (2003). *What's wrong with the Rorschach?: Science confronts the controversial inkblot test.* San Francisco, CA, US: Jossey-Bass.

Appendix: Sample items

Sample items from the *UPP* test	
Will to cooperate	I have been unfairly treated at work
Creativity	I often find a key to solving problems in the most unexpected places
Emotional intelligence	I can avoid doing reckless things when I'm upset
Social ability	I can easily make contact with new acquaintances
Positive attitude	I have not regretted choosing the profession I am in for one minute
Endurance	When I fail, I try again
Self-confidence	I see difficulties as challenges
Perfectionism	If you have decided to do something you should do it perfectly
Social security	To make contact with a stranger is not a problem for me
Narcissism	I am a member of the small group that can advance and be of great importance in society
Independence	What others think about me does not usually bother me that much
Extraversion	In a group, I am the person who takes responsibility
Emotional stability	I do not make mistakes, even if there is a mad rush to complete a task
Openness	I prefer concrete tasks rather than having to familiarize myself with abstract ideas
Agreeableness	I have many really good friends
Conscientiousness	Order is the basis for my work
Willingness to work	When I am at work, time moves at a snail's pace
Work interest	I often get tasks at work that strongly interest me
Job satisfaction	I feel satisfied with the work I do
Results orientation	I definitely want my work to yield results of significance
Willingness to work with changes	In most workplaces, I think better results can be achieved, if one is willing to participate in change
Control orientation	In my experience, it has been important to closely monitor employees’ and colleagues' work
Economic motivation	I think a lot about how I should be able to earn more money
Work/life balance	Things that I want to do at home tend not to be done, because of the demands of my job