



PSYKOLOGISK METOD AB



A third generation personality test

Lennart Sjöberg
Rapport 2010:3

Psykologisk Metod L Sjöberg AB arbetar med utveckling och användning av psykologiska test samt undersökningar av attityder och riskuppfattningar, andra psykologiska utredningar och tillämpad forskning.

Vår affärsidé är att bedriva arbetet i nära anslutning till den aktuella forskningen inom psykologin.

Skrifter utgivna av Psykologisk Metod AB

Sjöberg, L. (2008). Bortom Big Five: Konstruktion och validering av ett personlighetstest. Rapport 2008:1.

Sjöberg, L., & Möller, K. (2009). Sociala arbetsfunktioner och personlighet. Rapport 2009:1.

Sjöberg, L. (2009). *UPP*-testet: Kriterierelaterad validitet. Rapport 2009:2.

Sjöberg, L. (2009). *UPP*-testet: Korrektion för skönmålning. Rapport 2009:3.

Sjöberg, L. (2010). *UPP*-testet: Tredje generationens personlighetstest. Rapport 2010: 1.

Sjöberg, L. (2010). *UPP*-testet: Reviderad manual.

Sjöberg, L. (2010). *UPP*-testet: Mångfald gynnas av korrektion för skönmålning. Rapport 2010:2.

Sjöberg, L. (2010). A third generation personality test. Rapport 2010:2.

Abstract

The development of personality testing in the workplace has undergone three phases. The first generation of tests, such as Cattell's 16 PF and the British test OPQ, was characterized by complex systems for the description of the personality. These systems were simplified in part by the following generation of the test, which was based on the five factor model but that model was simple only at the horizontal 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 the third generation of personality tests. The test features focusing on "narrow" and work-relevant traits, inclusion of a few 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, measurement of current mood at the time of testing which can give another "warning signal", measurement of attitude towards the test ("face validity"), two types of narrative reports both to the person taking the test and the recruiter/psychologist – one based on normative comparisons and the other on ipsative (within-person) comparisons, measures of work related attitudes which are of value in themselves but can also be used as proxy criteria, greatly facilitating validation work.

Key words: personality test, impression management, mood

First generation: 1920-1990

Self-report tests of personality go back to the First World War. Until the 1960's the American tradition was represented by factor tests such as Cattell's 16pf - 16 factors - or Eysenck's much simpler MPI, which only had 3 factors. The American tradition, however, came to dominate. Complex tests were most common. But Cattell, creator of the 6pf, based his test on factor analysis of small data sets. The results have never been possible to replicate (Cooper, 2002). That the first generation of personality tests resulted in a very complicated test structures was therefore probably in part - maybe entirely - due to the fact that there was no access to effective statistical and psychometric methodologies.

Some older British tests are very often used in Sweden: PPA and OPQ. 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 3 statements in each block. The reliability of the 32 scales is doubtful. A meta-analysis of British research on the OPQ (Robertson & Kinder, 1993) showed that the test had a predictive 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 (Brown & Bartram, 2009).

The practical application of tests is changing only slowly. The most widely used test in Sweden, PPA, is an adjective list with roots in the 1920s (Marston, 1989/1928). It measures possibly some important dimensions, but far from all of potential importance in industrial psychology applications. Another old test is MBTI is based on Jung's personality theory from the early 1920s. See a critical discussion of MBTI elsewhere (Sjöberg, 2005) or Paul's lucid review (Paul, 2004).

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, around 0.3.

0.3 is a number that Mischel draw important conclusions from in his classic book from 1968 (Mischel, 1968), but already in 1921 the brothers Allport reported data at the same low level (Allport & Allport, 1921). They speculated then that it would be possible to obtain much stronger results in better tests. It has become apparent that they were wrong.

One of the basic problems of the first generation of personality tests was that it 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 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 that come into play. We do not even know how much of behavior, or job success, which in principle can be predicted on the basis of individual factors - the question is rarely discussed.

Second generation: 1990 - 2010

Around 1990 the notion of the “Big Five” was suggested (McCrae & Costa, 1987). It seemed to be a very attractive notion to apply just five factors to describe personality. American tests such as NEO-PI-R (Costa & McCrae, 1992) and HPI (Hogan, 1992) are based on this model. The international database IPIP has released test items, which may be translated and used. It also gives measures of the overall Big Five personality dimensions. 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. But complexity is still left in the form of sub-scales, so-called 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.

Second-generation Big Five tests have had a strong impact, especially in basic research. 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 now been made of the Big Five dimensions regarding their practical value in the workplace. The results have been disappointing, however (Morgeson, et al., 2007a, 2007b). Big Five dimensions have not produced an improved predictive power.

Second generation tests also did not solve the problem of impression management (IM) in high-stakes testing.

The third generation

The first and second generations of personality tests were facing problems that they did not manage to solve and which can be managed effectively only in a new generation of significantly enhanced tests. I shall now deal with these problems and start with impression management. My examples come from the *UPP*-test. *UPP* stands for Understanding Human Potential.

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 test are very 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 too much to handle cognitively and doubts arise if all of these scales can indeed be measured reliably with a test period of 30-40 minutes.

The UPP test

UPP is a personality test intended for applications in industrial and organization settings. The test has been developed by Professor Lennart Sjöberg at the Stockholm School of Economics (SSE) and Psykologisk Metod AB, a consulting firm in applied psychology¹.

The test consists of 232 items. Most are self-report items but some are of the performance type. It takes about 40 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.

There is a norm group consisting of about 1200 persons. All scales are reliably measured and have been concept validated. External criteria have also been used with success (manager career). It is very hard to fake test results on *UPP* because impression management (IM) scales are used to correct for attempts to exaggerate the results.

The test is continuously improved in further research.

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

- Social ability
- Emotional intelligence
- Will to cooperate
- Endurance
- Positive attitude
- Self-confidence
- Creativity
- Perfectionism

It also measures the usual Big Five personality dimensions:

- Emotional stability
- Extraversion
- Conscientiousness
- Openness
- Agreeableness

These 13 (8+5) basic scales are used to construct aggregate measures of:

- Ego strength
- Stress sensitivity
- Managerial ability

¹ Sjöberg was a professor of Psychology at SSE 1988-2006. He was previously a professor at the University of Göteborg and Visiting Professor at Stanford University, and University of California, Berkeley. For further information, please see <http://www.dynam-it.com/lennart/>.

Aggregate scales can also be constructed for special applications to measure both the presence of desirable traits and the absence of undesirable traits.

UPP also 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
- Result orientation
- Work-life balance

These dimensions are 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.

Most of the scales use a self-report format. However, a special section includes the task of identifying emotions in facial expressions of the type depicted in Fig. 1.



Fig. 1. Types of facial expressions used in the *UPP* test.

The facial expressions are judged on 8 emotion scales; consensus is used to define the “correct” answer (Sjöberg, 2001, 2008b). Relations between self-report and performance scales of emotional intelligence are fairly weak, much like the situation in research on these topics (Joseph & Newman, 2010). However, both types of scales carry important information (Engelberg & Sjöberg, 2005). The concept has many interesting implications (Engelberg & Sjöberg, 2004; Engelberg & Sjöberg, 2006; Sjöberg & Littorin, 2005).

The test also creates a narrative report about each test taker, based on explicitly described principles.

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 8 focused scales (see above) make possible a dramatically improved validity when compared to the traditional Big Five, about 8 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.
- Each test taker is assessed normatively, in relation to a norm group, and intra-individually, in relation to his or her data only.
- The test assesses aggregate variables such as ego strength.
- 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 available in 7 modules:

- A. The complete test
- B. Personality scales only (8+5 scales)
- C. Screeningtest. Five of the scales are used for a quick and efficient screening of many applicants.
- D. Adjustment to the current work situation, 6 scales.
- E. Social and emotional ability.
- F. Ego strength, short scale.
- G. Traditional Big Five dimensions.

All modules include our method for correcting for IM. In the following sections we give selected results for some of the advantages of the *UPP* test.

Correcting for IM in the UPP test

A social desirability scale, or lie scale, can be used in a selection situation for weeding out those that have high value, but that strategy does not imply that the effect of impression management disappears, it is only slightly mitigated (Sjöberg, 2008a). 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 1 gives an example from a study of a large data set² from job seekers who had taken the UPP test (the screening module) in connection with applying for a job (Sjöberg, 2009). Correlations be-

² 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>.

tween personality variables and a measure of impression management are apparently quite variable. Hence, a global approach towards correcting for IM will fail.

Table 1. 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

In the *UPP* test, a regression model is fitted to each test variable separately and residuals are used to estimate corrected scale values. Fig. 2 shows the results on the same data set as Table 1. 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.

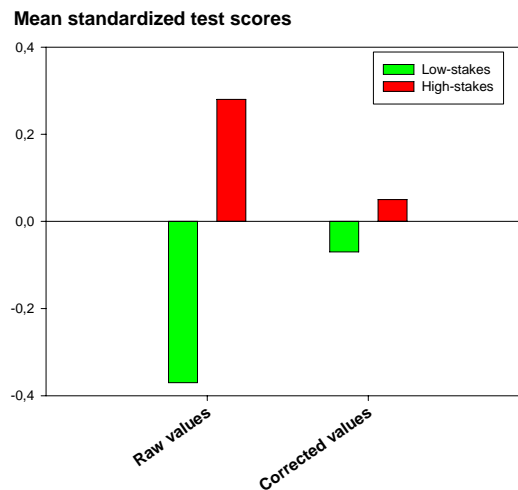


Figure 2. Mean values of personality dimensions before and after correction for IM, high-stakes and norm data.

Fig. 3 gives corresponding results from an experimental study where some participants were instructed to fake good, others just to answer in an honest manner.

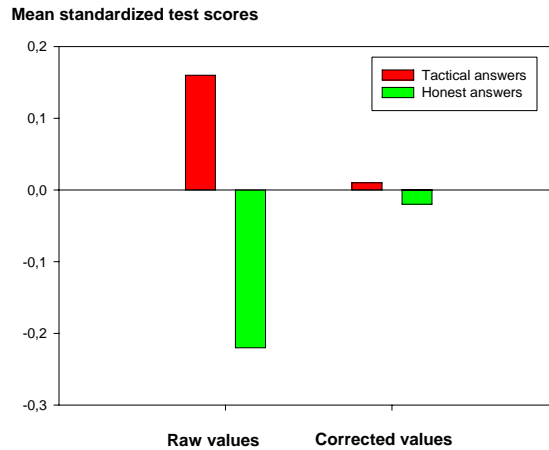


Figure 3. Mean values of personality dimensions before and after correction for IM, experimental study.

It is interesting to note that gender differences in applications for management jobs, favouring men over women, are greatly reduced due to IM correction, see Fig. 4.

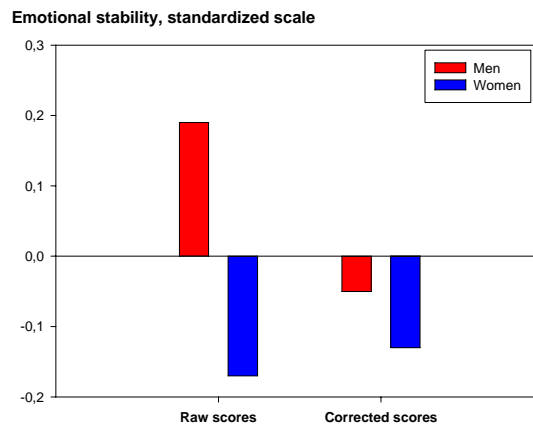


Figure 4. Raw and corrected scores in emotional stability of men and women applying for management jobs,

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.

Mood and test results

It is likely that responses on a personality test are affected by the current mood state of the testee, but this is seldom attended to in practical testing. Table 2 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 (Sjöberg, Svensson, & Persson, 1979),

which has been widely used. The data are from testing applicants to the Stockholm School of Economics.

Table 2. Correlations between Big Five personality scales and current mood state, high-stakes testing, N=210.

Personality scale	Mood dimension			Pooled measure of mood
	Happy-sad	Alert-tired	Calm-tense	
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. 5, where the 10 percent worst mood cases are compared with the rest of the testees in standardized scores.

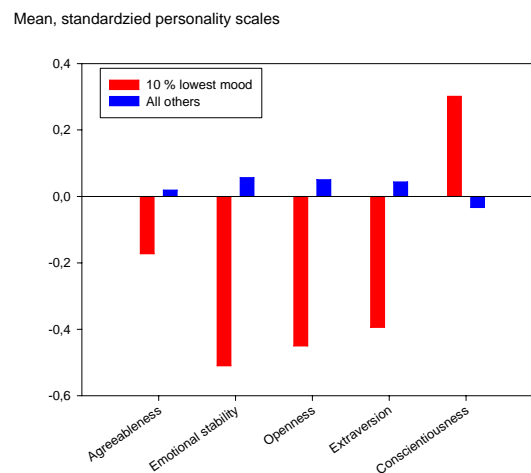


Figure 5. Mean personality scales values for the 10% of the test takers who were in the worst mood, compared to all others. Data from a high-stakes situation.

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 subsequent 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 scales, 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. 6 shows the relationship between mean data quality and level of education, separately for men and women.

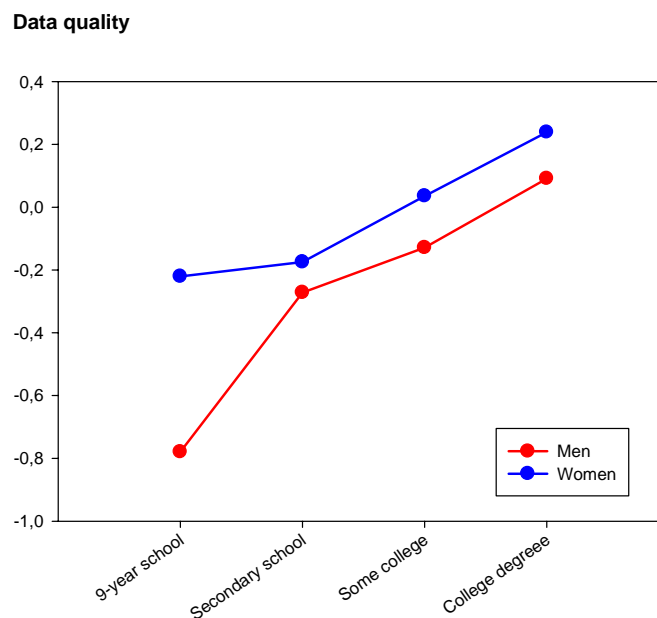


Figure 6. Mean data quality for men and women at different levels of education.

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 . 5.

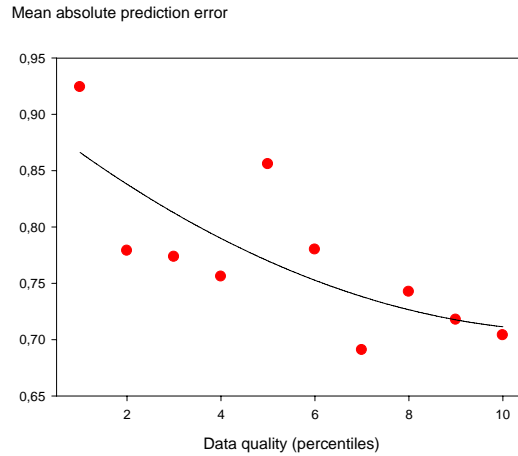


Figure 7. Mean absolute prediction error plotted against data quality (percentile groups).

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

Attitude towards the test: “Face validity”

It is important to know about the test takers’ attitudes towards 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 8 questions measuring attitude towards the test; these questions are correlated and are used to estimate a pooled measure of attitude. The distribution of attitude ratings for 103 test takers is provided in Fig. 8.

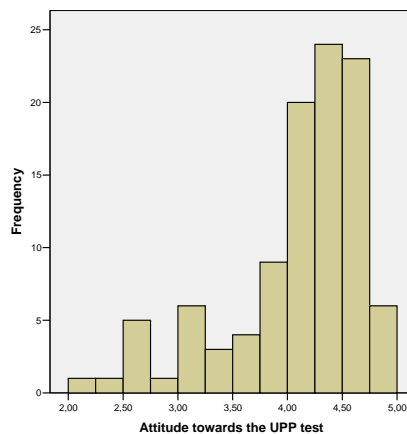


Figure 8. Distribution of mean ratings of attitudes towards the UPP test.

The figure shows an overwhelmingly positive attitude towards the test.

Validation of the UPP test

External criteria

Social job skills were measured in one study and related to the UPP, see Table 3.

Table 3. 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. A gender difference in emotional intelligence is illustrated in Fig. 9. It is well known that women tend to have a higher emotional intelligence than men (Joseph & Newman, 2010).

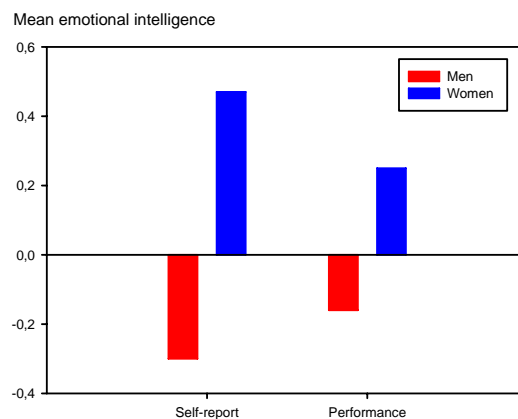


Figure 9. Gender and emotional intelligence

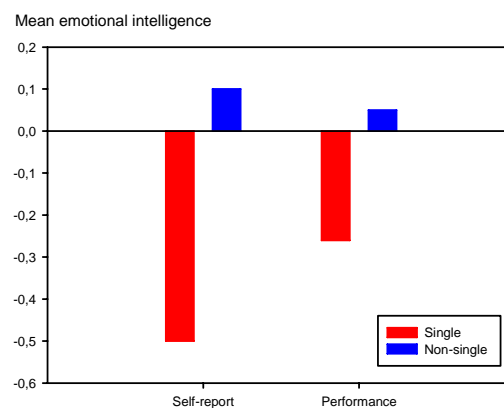


Figure 10. Civil status and emotional intelligence.

Management career was related to the *UPP* test, see Table 4. The criterion was binary: manager or not manager.

Table 4. Hierarchical binary regression analysis of manager responsibility against the *UPP* test. N=107.

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

The criterion validity of the *UPP* is very satisfactory in these data. A final validation of the test is reported here, using proxy variables. See Table 5.

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

Criterion	The FFM model	The <i>UPP</i> 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 in this paper and much better than the common results for tests of the first and second generations, which tend to lie in the interval 0.2 – 0.3.

Conclusion

Personality testing has developed slowly over almost a century. Some still widely used tests (first generation, 1920-1990) are based on notions from the 1940's or even earlier. In the second generation (1990-2010) of tests, Big Five dimensions dominated, but they are too broad to be of practical value. The *UPP* test solved these and a number of other problems. The original Swedish version is now being translated and standardized for a number of other languages.

References

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

- Brown, A., & Bartram, D. (2009). *Development and psychometric properties of the OPQ32r. Supplement to the OPQ 32 technical manual*: SHL.
- Cooper, C. (2002). *Individual differences. 2nd edition*. London: Hodder Education.
- Costa, P. T., Jr, & McCrae, R. R. (1992). *NEO PI-R professional manual*. Odessa, FL: Psychological Assessment Resources, Inc.
- Engelberg, E., & Sjöberg, L. (2004). Internet use, social skills, and adjustment. *Cyberpsychology & Behavior*, 7(1), 41-47.
- 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.
- Engelberg, E., & Sjöberg, L. (2006). Money attitudes and emotional intelligence. *Journal of Applied Social Psychology*, 36(8), 2027-2047.
- Hogan, R. (1992). Hogan Personality Inventory. *Psychological Test Bulletin*, 5(2), 130-136.
- 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(1), 54-78.
- Marston, W. M. (1989/1928). *Emotions of normal people*. Ormskirk, Lancs.: Thomas Lyster.
- 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.
- Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology*, 60(4), 1029-1049.
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, 60(3), 683-729.
- 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.
- Robertson, I. T., & Kinder, A. (1993). Personality and job competences: The criterion-related validity of some personality variables. *Journal of Occupational & Organizational Psychology*, 66(3), 225-244.
- Sjöberg, L. (2001). Emotional intelligence: A psychometric analysis. *European Psychologist*, 6, 79-95.
- 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(1), 21-28.
- Sjöberg, L. (2008a). *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.
- Sjöberg, L. (2008b). 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.
- Sjöberg, L. (2009). *UPP-testet: Korrektion för skönmålning. (The UPP test: Correction for impression management)*. Stockholm: Psykologisk Metod AB.
- Sjöberg, L., & Littorin, P. (2005). Emotional intelligence, personality and sales performance. In K. B. S. Kumar (Ed.), *Emotional intelligence. Research insights* (pp. 126-142). Hyderabad, India: ICFAI University Press.

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