

The Swedish version of Thomas PPA: Critical comments

Lennart Sjöberg
Stockholm School of Economics

March 30, 2005

Thomas PPA is a commercially successful methodology widely used in Sweden, among other countries. It is used e.g. in employment interviews and for team-building. It has recently been assessed by STP (the Foundation of Applied Psychology) (Stiftelsen för tillämpad psykologi, 2004). This Swedish foundation is closely connected with the Swedish Psychological Association (SPA) and offers a service to test producers whereby they can have their tests evaluated. The procedure is closely similar to the evaluations carried out by the British Psychological Association (BPA). Several large corporations in Sweden now require all tests to have gone through this procedure, and so do some municipalities and the Government. The current issue of *Personal och Ledarskap*, a Swedish magazine widely read by HR professionals, contains a detailed overview of STP and the tests which have so far been assessed, among them PPA. There is also an interview with Per Falck, head of the Swedish agent of PPA, and comments on their expansion abroad, among them a very large program with IKEA in Russia.

The purpose of the present note is to summarize STP's assessment of the PPA, to offer some additional comments on theory and empirical results and to compare with the Predictive Index (PI).

There has been little Swedish research on the PPA, as far as I know none at all for many years, but a recent report by Irvine and Lindelöw may be noted (Irvine & Lindelöw-Danielsson, 2004). They obtained reliability and validity data and some positive results for the PPA. It is my impression that the very existence of STP and its assessment program has been an important factor in stimulating Swedish research on test properties, certainly for PPA, but also in other cases such as the MBTI.

Conceptual basis

Thomas PPA is an instrument¹ for describing personality in four main dimensions:

Dominance
Inducement
Submission
Compliance

The PPA is a DISC instrument, after the initials of the four basic dimensions.

¹ . It may be debated whether the "system" is a test or not. However, its purpose is the same as that of other personality tests, i.e. to describe the test taker's personality, so I will refer to it as a test.

The first two dimensions refer to active behaviour, the last to a passive, wait-and-see attitude. Dominance and compliance refer to behaviour in an antagonistic setting. Inducement and submission refer to friendly settings. These factors were originally suggested by Marston (Marston, 1989/1928) in 1928. The “Thomas System” builds on the work by Thomas Hendrickson, starting in the 1950’s (Hendrickson, 1983). Irvine has recently given a thorough description of the conceptual framework, as well as a review of the empirical properties of the test (Irvine, 2003). Hendrickson’s original and basic work was never published and is no longer available since it was destroyed in a fire many years ago. Be that as it may, Thomas Hendrickson was apparently not a very active researcher. In the comprehensive data base PsycLit, which lists most of scientific psychology publications from 1840 and onwards, he is not mentioned even once. Thomas International has published two technical reports by him, one from 1983 and the other undated.

In the PPA, 96 adjectives are used, organized in 24 groups of four. The test taker is instructed to check, in each group, the adjective which best describes her, or is worst fitting as a description of her. Each adjective belongs to one of the four dimensions. Hence, one may derive a profile for “best” descriptors and one for “worst” descriptors. The former is asserted to be a description of how the test taker reacts in the work environment. The latter is asserted to be a description of how she reacts under stress. The difference between the two profiles is assumed to depict her self-image. The notion that the profile of “best” descriptors provides a (partial, see below) picture of the generalized self image is reasonable but it is unclear how the other two profiles (“worst” and the difference between “best” and “worst”) can be interpreted the way they are. STP points to this weakness of the system. They also say that there is a lack of theoretical justification for the interpretations made on the basis of a PPA, and that they seem to be “incomprehensible” and “illogical”. These are strong criticisms indeed. They also argue that the PPA is hard to use in practice, and that other methods are available, using simpler and straightforward adjective check lists. This is also a very serious criticism.

Historically speaking, Marston’s basic work is obsolete and obscure. Hendrickson’s contribution seems to have been to bring it into the mainstream psychology of the 1950’s. At that time, Allport was a leading name in personality psychology and Thurstone had published factor analytic work (Thurstone, 1951), soon to be followed by a host of studies of such famous psychologists as Cattell, who devised the 16pf test, and Guilford. The field produced more and more factors and was increasingly confusing until McCrae and Costa (McCrae & Costa, 1987) reported in 1987 that a simple five-factor model (the FFM model or “the big five”) was sufficient to account for self-reported personality scales and dimensions. The FFM is now a bench-mark standard. The Thomas PPA matches some of it, but not all. It is therefore, at best, an incomplete system for describing self-reported personality.

The FFM is a very important and fairly recent innovation in testing, a second one being emotional intelligence (Law, Wong, & Song, 2004; Sjöberg & Engelberg, 2004). Made popular by Goleman some ten years ago (Goleman, 1995), emotional intelligence testing is now much in demand and research has reached the level of some 150 research reports published every year. Before 1990, the concept was virtually unknown.

A third innovation is risk perception and risk taking behaviour (Law et al., 2004). These dimensions are so far less attended to, but there is very clear evidence for the importance of perceived job related risk to, e.g., work motivation (Sjöberg, in press-b).

None of these innovations in testing and personality psychology are considered in PPA which sticks to its roots from the 1950's. This is particularly noteworthy if the PPA is used as a stand-alone test, i.e. if no other tests are used besides it. The PPA is indeed often used as a stand-alone test. It is likely that other consultants have a similar strategy. Even if they don't use the PPA, they have only one personality test among the methods they employ in their work. Typical cases would be MBTI and the OPQ. Further information about the Swedish situation when it comes to personality tests can be found in a report (Sjöberg, 2004b).

Is personality complex or can it be described for all or most practical purposes on the basis of a simple test, which is administered in 10 minutes and can be scored by a computer? Most psychologists would probably be sceptical about such a simplistic approach. The predictive value of a more complex approach has been well established in current meta-analytic work (Salgado, 1997). It should also be noted that explicit reference to the work situation provides an important practical improvement in personality testing (Bing, Whanger, Davison, & VanHook, 2004). Traditional personality testing was based on the assumption of a generalized personality, long ago found to be inadequate (Mischel, 1968).

Psychometrics

The PPA yields ipsative scores² since it is based on comparing adjectives to each other. In other words, they could all yield a bad fit, or a good fit, but the tester would never notice. In turn, this means that very serious statistical problems arise when the four dimensions are combined for predictive purposes. The PI does not have this problem since it is basically normative, not ipsative.

Reliabilities of the PPA are found by STP to be acceptable, with the exception of homogeneity coefficients. The latter lie at the level 0.6 which is far too low for individual diagnosis.

Strangely enough, STP does not discuss the report by Danielsson (Lindelöw Danielsson, 2002) in detail. She reports the following "internal consistency reliabilities", in her Table 2, p. 7 of her report (translated here), see Table 1.

². There are some thorny issues connected with ipsative scores: what they are and how they can be used. Yet, the basic idea is a simple one. People make choices or comparative judgments and comparisons can then only be made *within* the individual, never *between* individuals. With such data, we obviously cannot say that one person is, for example, more dominant than another. A person with a very high dominance score can be quite timid if she is even lower in other aspects of personality with which she has compared dominance adjectives. Conclusion: ipsative scores cannot be used in selection. PPA can for *logical* reasons not be used in selection. Yet, it is so used.

Profile I	D	0.6981 ¹
	I	0.6374
	S	0.6650
	C	0.5078
Profile II	D	0.7328
	I	0.5973
	S	0.5680
	C	0.5091

Note 1. It is unclear why four decimal places are reported.

These results hardly match the text of her report. The 8 coefficients of her Table 2 (our Table 1 above) are said to be 12, and lie between 0.7 and 0.9 with two exceptions. Not true. As can be seen they are much lower, between 0.51 and 0.73 with a mean of 0.62. In particular, 3 out of 4 values for Profile II are lower than 0.6! These values are clearly unacceptably low. Since STP reports that these values lie in the interval 0.6 – 0.7 they cannot have read her Table 2, but only her summary of the table, which is erroneous.

It should also be noted that a profile interpretation is based on *differences* between scale scores, and differences are even less reliable than the primary scores. Hence the low reliability of the PPA in its Swedish version is a very serious drawback. Unfortunately, STP does not notice the additional problem with difference scores and profile interpretation.

The PI fares much better when homogeneities of the basic dimensions are estimated (Sjöberg, 2000, 2003). The PI reliabilities are around 0.8 in the Swedish version, a sufficiently high value.

Criterion validities are found to be around 0.3. This is fairly typical for a personality test, but it should be noted that considerably better results are often achieved with a set of dimensions, e.g. the FFM model dimensions. Explained variance is only about 10 percent with the PPA. However, all these results are of doubtful value due to the ipsative scoring used. The PI seems to have a similar level of criterion validity but it is not questionable to the same extent as PPA since it uses normative scoring.

It should be noted that the PPA has a very wide range of applications, according to its agent. Little or no research seems to cover such applications which go far beyond simple prediction of criteria in the classical sense. Take team-building as an example. There is no knowledge about the PPA's efficiency in this respect. A long list of similar applications could be made. STP does not stress or even note the need for scientific support for these applications.

The construct validity is based on empirical comparisons with other tests. There has been some limited success in this respect with the PPA (a similar situation with the PI), but STP complains that there has been no attempt to compare the PPA with other DISC instruments. STP is otherwise generous in its judgment of the construct validity results. They accept information which is partly restricted to statistical significance. However, as is well known,

significance is largely a matter of sample size and effects size or correlation is essential information when the validity of an instrument is to be assessed³.

It would be interesting to compare two DISC instruments such as the PPA and PI. As will be shown below, there is reason to believe that they are *not* equally good.

Norm data are said by STP to be acceptable. This is not surprising, given the extensive applications of the PPA.

It can be added that STP does not inquire into the factorial structure of the PPA. This problem has been thoroughly analyzed with SEM models with regard to the PI (Sjöberg, 2003) which is found to give a reasonable fit to the hypothesized structure. The ipsative nature of the PPA makes it very difficult to analyze the factor structure of PPA data. However, the adjectives used by the PPA could be studied and analyzed in a PI design which gives normative data, or preferably on the basis of ratings in a Likert scale format (category scales with several steps). Such a study could easily be carried out.

Also, STP does not discuss gender differences, or differences between ethnic groups. These are increasingly important matters, where it is known that PI has desirable properties.

The STP could also have been more penetrating when it comes to social desirability bias. True, the ipsative format should make the PPA less amenable to such bias, but there is no guarantee that it is wholly absent and the question needs empirical study. The PI has been studied in this regard and it was found that bias was fairly modest (Sjöberg, 2003).

Development of the PPA

STP points out that description of how the test was developed are incomplete. For example, how were the English words translated? It is well known that such translations are quite difficult. They are of essential importance in an adjective test such as the present one. PI has a high level of ambition in this respect and has relatively recently greatly improved the quality of its translation, using back-translations and extensive bi-lingual discussions.

Layout, manual, instructions

STP criticizes the PPA material for being incomplete and in some respects hard to understand and use. It is noteworthy that the PPA is sold to non-professionals and that the British agent even refers to it as an advantage that the user need not have an education in psychology. Since this is their policy, it would seem that very clear and complete information is essential.

Summing up

The STP assessment of PPA is quite critical. On the positive side, they note what they judge as acceptable levels of reliability and criterion validity. Construct validity is found to be low and there is a lack of research on this topic. The most important criticism concerns the conceptual and theoretical basis of the test, which is found to be obscure (and possibly obsolete). The advantages of using PPA rather than simple adjective lists of a more straightforward nature are found to be lacking.

³ It may be noted that STP has made the same mistake in its assessment of the Myers-Briggs Type Indicator, see my critical disquisition of STP and this test (Sjöberg, 2004a, in press-a).

Since the PPA has a large share of the personality test market in Sweden, one wonders how this can be the case, given the negative highly critical picture of the STP assessment and also the additional critical remarks I have made above. It can only be concluded that the market is one thing, the scientific basis another. Paul has recently described the markets of several of the most well known personality tests (Paul, 2004). The technical and scientific basis of a test seems to be relatively little important for commercial success.

The STP report can undoubtedly be used for selective citations, just as was done with MBTI by their Swedish Agent (Psykologiförlaget and their owner, Hunter Mabon). A marketing argument is sure to be that a test has been assessed by the STP, regardless of what was written in the STP report. (Who will read it?)

A comparison between the PPA and the PI is pertinent. The two instruments have some common historical roots in the work of Marston. They are also similar in working with few basic concepts, four dimensions⁴. Having noted that, the following points should be stressed:

- PI has higher reliability values
- PI has been found to fit reasonably well to the four-factor model. There is no Swedish research on the PPA pertinent to this issue
- PI is not gender or age biased, nor is it biased with regard to ethnic background
- PI has criterion validities at the same level as the PPA
- PI yields normative rather than ipsative scores, an advantage because it is the possible to compare individuals. How can you make, e.g., selection decisions without being able to compare individuals? It is also a great advantage in the sense that statistical analyses are straightforward and not marred with psychometric difficulties, as in the PPA case with its ipsative scores.
- Social desirability bias is not taken into account in the PPA; for the PI it has been found to be modest

PPA and PI are compared in Table 2 below.

⁴ . It is a good idea to work with simple but powerful basic dimensions. Compare with the OPQ, 32 dimensions, and Cattell's 16pf, 16 dimensions. How can you make sense out of 32 dimensions? Remarkably, the OPQ is another success story in spite of very meagre scientific support and a validity level lower than most self-report personality tests (around 0.2).

Table 2. Comparison between PPA and PI, Swedish versions		
Aspect	PI	PPA
Sufficient reliability (alpha coefficient)?	Yes	No
Sufficient stability over time?	Yes, US data	Yes
Four factor structure established?	Yes	No
Quality assured translation of crucial terms?	Yes	Not known
Criterion validity at the 0.3 level (normal level)?	Yes	Yes, subject to interpretation difficulties (ipsative format, see text)
Gender, age and ethnic bias?	No	Not known
Social desirability bias?	Modest	Not known
Can be used for selection?	Yes, it is normative	No, it is ipsative

Comments on STP assessment

The STP process is new in Sweden, and so far only a handful of tests have been analyzed. As noted in the introduction, there is a trend for some large customers of tests and testings to demand STP assessment. We do not yet know if this will decrease demand for non-assessed tests but it seems likely.

It should be noted that STP does not publish definite conclusions such as “approved” or “not approved”. They will typically rate a test in a number of dimensions, some of which are less essential for the value of the test. Some test producers and agents have utilized this circumstance in their marketing. As an example, the MBTI was given a highly critical assessment when it came to essential properties, a more positive one in several peripheral aspects. The test agent computed a mean assessment and argued that the test came out in a very positive manner from the process. People involved in the process probably did not like this, but what could they do? It should also be noted that the STP reports tend to be quite technical, they are hard to get a copy of (they are sold at a rather high price by STP which retains copyright). It is questionable if practitioners can and will read those reports.

The STP process is modelled on the one designed by the BPA. Swedish circumstances do not quite fit. For example, large sample sizes of 100 and more are required for validity studies. It is not obvious that a well designed validity study with, say, 30-50 participants, cannot be quite informative.

The STP process concentrates on the test as such. It does not really discuss, as I have done above for PPA, how well the test is connected with scientific development of the field, nor its

role in a practical context. Does the PPA efficiently promote team building? Who knows? How does it work in a context together with other tests? Can it contribute to predicted variance in a setting of several tests, including the FFM factors and emotional intelligence? These are all pertinent questions but the STP disregards them concentrating on the test as such. This may be reasonable but one should be aware of the limitations. There are many questions about a test not answered or attended to by STP.

Ethical questions are a set of concerns not taken up by STP. I have discussed gender and ethnic background above. There are difficult questions having to do with self-knowledge. Do tests such as the PPA or MBTI promote self-knowledge? Is it even desirable that they do? Assertions are often made that such tests promote self-knowledge but research on these topics has hardly begun.

The STP process thus has many limitations. One could add that it has been criticized for using anonymous reviewers. BPS or the Buros Institute⁵ name their reviewers. The fee for an STP assessment is also substantial (SEK 50,000).

⁵ . An excellent on-line American service which reviews hundreds or thousands of tests. Reviews can be downloaded for US \$ 15. Reviews appear to be objective, fair and competent.

References

- Bing, M. N., Whanger, J. C., Davison, H. K., & VanHook, J. B. (2004). Incremental Validity of the Frame-of-Reference Effect in Personality Scale Scores: A Replication and Extension. *Journal of Applied Psychology*, 89(1), 150-157.
- Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.
- Hendrickson, T. M. (1983). *Personal profile analysis: a technical manual*. Marlow: Thomas International Systems (Europe) Ltd.
- Irvine, S. H. (2003). *Making people profitable. Personal profile analysis. The technical resource book*. Marlow: Thomas International Ltd.
- Irvine, S. H., & Lindelöw-Danielsson, M. (2004). *PPA validity studies in Scandinavia: An occupational validity study in Sweden*. Stockholm: SLG International.
- Law, K. S., Wong, C. S., & Song, L. J. (2004). The construct and criterion validity of emotional intelligence and its potential utility for management studies. *Journal of Applied Psychology*, 89(3), 483-496.
- Lindelöw Danielsson, M. (2002). *PPA: En svensk reliabilitetsstudie*.
- 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.
- 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.
- Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology*, 82, 30-43.
- Sjöberg, L. (2000). *The psychometric structure of the Swedish version of the Predictive Index (PI)*. Stockholm: Economic Psychology Unit.
- Sjöberg, L. (2003). *Properties of the new Swedish version of the Predictive Index*. Stockholm: Economic Psychology Unit, Stockholm School of Economics.
- Sjöberg, L. (2004a). Myers-Briggs-testet - en sällskapslek. (The Myers-Briggs test - a parlour game). *Psykologtidningen*, 50(13/04), 12-14.
- Sjöberg, L. (2004b). Personlighetstest i arbetslivet: historik och aktuell forskning. (Personality tests in industry: History and current research). In G. Sevón & L. Sjöberg (Eds.), *Emotioner och värderingar i näringslivet* (pp. 171-229). Stockholm: Ekonomiska Forskningsinstitutet - EFI.
- Sjöberg, L. (in press-a). En kritisk diskussion av Myers-Briggs testet. (A critical discussion of the Myers-Briggs test). *Organisational Theory & Practice*.
- Sjöberg, L. (in press-b). Risk and willingness to work. *International Journal of Risk Assessment and Management*.
- Sjöberg, L., & Engelberg, E. (2004). Emotionell intelligens: Teori och empirisk forskning. (Emotional intelligence: Theory and empirical research). In G. Sevón & L. Sjöberg (Eds.), *Emotioner och värderingar i näringslivet* (pp. 65-116). Stockholm: Ekonomiska Forskningsinstitutet - EFI.
- Stiftelsen för tillämpad psykologi. (2004). *Granskningsrapport - Thomas Systemet Personlig Profil Analys, Thomas PPA*. Stockholm: Stiftelsen för Tillämpad Psykologi.
- Thurstone, L. L. (1951). The dimensions of temperament. *Psychometrika*, 16, 11-20.

