Project Epsom: How Valid is Your Questionnaire?

Phase 1: Saville Consulting Wave®, OPQ®, Hogan Personality Inventory & Development Survey, 16PF5, NEO, Thomas International DISC, MBTI and Saville Personality Questionnaire Compared in Predicting Job Performance

This research paper contains content originally presented in Professor Peter Saville's Keynote Speeches at:

The British Psychological Society Division of Occupational Psychology Conference, Stratford-upon-Avon, January 2008: "Personality Questionnaires – Valid Inferences, False Prophecies"

The Psychological Society of South Africa Annual Conference, Johannesburg, August 2008: "Does Your Test Work?"

A&DC Conference, Institute of Directors, London, November 2008: "A Comparison of Leadership in Business and Elite Athletes"

The Authors

Professor Peter Saville BA, MPhil, PhD, FBPsS, C.Psychol, FRSA was Chief Psychologist at the Test Division of the National Foundation for Educational Research by the age of 27. He was Founder and Chairman of SHL and created the OPQ®. In 2001 Peter was voted as one of Britain's top ten psychologists, the only industrial psychologist listed. In 1998 he was nominated as one of the UK's top entrepreneurs. His portrait hung in the National Gallery, London after being awarded the British Psychological Society Centenary Award for Distinguished Contributions to Professional Psychology. He has published over 300 papers, books, questionnaires and keynote speeches, as well as having been interviewed on radio and television around the world. Peter has consulted with many of the FTSE and Fortune 100 companies, as well as public bodies and the United Nations. Now International Chairman at Saville Consulting, he continues to research valid and improved ways to measure workplace performance.

Rab MacIver BSc, MSc, C.Psychol, previously worked at SHL on the revised version of the OPQ®. Now a director at Saville Consulting whose primary role has been in the research and development of the Wave portfolio of products, Rab is a commercial, client-focused psychologist with strong technical development skills and a proven track record in formulating and implementing effective, research based,recruitment and development solutions, both in private and public sector organizations.

Dr Rainer Kurz BSc, PhD, C.Psychol is a director involved in the key research and assessment technology programmes at Saville Consulting. He led the way in the development of the Saville Consulting aptitude assessments, expert systems and competency frameworks that are now used in more than thirty countries.

Tom Hopton BA Oxon is a consultant at Saville Consulting who graduated from Oxford University as an Experimental Psychologist. Working in the research and development of various psychometric instruments, he has coordinated a major validation study of personality questionnaires. A swimmer once ranked second in the British Isles, Tom is currently writing a book on the personality of elite sportspeople, entrepreneurs and other well-known individuals.

Acknowledgements

We would like to thank the large number of external reviewers who have made this paper possible. We would also like to extend our thanks to those people who were involved in the data analysis and administration of this extensive project, including Katie Herridge, Anna Mitchener, Gail Moors, Celine Rojon, Jemaine Saville, Hannah Staddon and Karen Tonks. Our sincere thanks also go to Ian Woosnam OBE for giving permission to show part of his Wave profile.

Project Epsom: How Valid Is Your Questionnaire?

Abstract

A major research initiative, Project Epsom, compared the validities of a range of the most popular personality questionnaires using the same sample and the same work performance measures. In this study the Saville Consulting Wave® Professional Styles was the most valid questionnaire in terms of measuring job performance. The questionnaires compared were validated against the externally-developed SHL Great Eight competency framework (Kurz & Bartram, 2002) and a global performance measure, in order to ensure fairness of comparison and to avoid bias towards the Saville Consulting questionnaires. Great care was taken in the use of these work performance criteria and the equations for predicting work performance published by Bartram (2005) were utilized for the Occupational Personality Questionnaire (OPQ®).

The questionnaires were also compared to other models of work performance, including the extensive Saville Consulting model of work effectiveness (Kurz et al., 2009). Against this model, the Saville Consulting questionnaires performed better still, but for the purposes of this paper these results are not presented here. The Saville Consulting Wave Professional Styles questionnaire therefore outperformed the OPQ® against its own model of work effectiveness.

The newly-developed Saville Personality Questionnaire (Saville PQ™; Saville et al., 2009a) also performed as well, if not better than the OPQ® and many other established questionnaires. The Saville PQ was developed using the same approach as the OPQ®, takes under 15 minutes to gather both normative and ipsative responses and makes a crucial distinction between a person's talents and motivations. Many of the other questionnaires compared in Project Epsom did show at least a moderate level of validity in measuring job performance.

In considering the results from this research, the present paper also provides an initial orientation in the key concepts surrounding personality questionnaires and offers readers guidance on how to select the most appropriate questionnaire for measuring work performance. This paper finally considers why the Saville Consulting questionnaires were found to be the most valid measures of work performance.

Background

Nelson Mandela once asked "does anybody really think that they didn't get what they had because they didn't have the talent or the strength or the endurance or the commitment?" In this statement, Mandela recognizes the importance that personality plays in driving success in life. For example, a representative reported that in one major office technology company, some 80% of their sales consistently came from just 20% of their best salespeople.

What, then, is personality? There has been no shortage of answers to this question. In developing the OPQ® Saville et al. (1984) defined personality as "an individual's typical or preferred ways of behaving, thinking and feeling". A similar definition has been proposed by Costa and McCrae (1992) with their Big Five model of personality. More recently, Digman (1997) distinguished between Alpha personality characteristics and Beta personality characteristics, a distinction which is similar to that between people who "get along" and those who "get ahead".

Cronbach (1970) saw personality as a "behavioral posture" and, as with other researchers, Cattell (1965) emphasised the criticality of validity when he stated that personality is "that which enables us to predict what a person will do in real-life situations". In our application, validity represents job success.

It is increasingly acknowledged in the contemporary world that job-relevant and well-constructed personality questionnaires can be used successfully to measure what a person will do in real-life situations, and in particular to improve decisions in the selection and development of people at work. There is a proliferation of personality questionnaires available purporting to offer the means to achieve this. In the field of personality assessment, there are a number of reasons why it can be difficult to choose between the different questionnaires available and to select that which is most suitable.

Some test publishers use complicated jargon which may confuse many test users, while others refrain from publishing negative findings. Statistical techniques can be misapplied in an attempt to overestimate the effectiveness or usefulness of a test. For example, statistical procedures might only be carried out on the

top and/or bottom 10% of people in the sample, ignoring the majority of the sample and vastly inflating the apparent relevance of the test. Additionally, some tests are merely compared with other tests to assess the degree to which they agree in their measurement. Such correlation techniques, however, do not ensure that the test necessarily demonstrates job-relevance or will measure performance at work. As Wiggins (1973) succinctly puts it:

"Regardless of the theoretical considerations which guide scale construction or the mathematical elegance of item-analytic procedures, the practical utility of a test must be assessed in terms of the number and magnitude of its correlations with non-test criterion measures"

The **validity** of a test in this context is the degree of relevance the test has in assessing effectiveness at work. A valid test must be able to measure how the test-taker is likely to perform in a given job. Data must be presented to back this up. If no evidence is presented to show that a test works, it should not, quite simply, be used to make decisions which could impact on people's careers and well-being at work. Choosing valid tests with established links to performance drives superior selection methods and in turn makes an organisation more effective by driving improved individual performance. Needless to say, validity is the single most important characteristic of any test and concerns whether a test actually works.

Other important concepts in testing include **norms, reliability** and **return on investment (utility)**. Test **norms** such as "percentiles" and "stens" show how an individual compares to a relevant sample of people. Norms are of course useful for such comparisons, but do not in themselves "prove" that a test works: they are not the sine qua non of testing. Indeed, there are occasions where one does not even need to have norms. For example, filling job vacancies by selecting the highest performers on a valid and job-relevant test can result in improved productivity, without necessarily comparing these scores against an external norm group. In this instance, the test could be highly valuable to the organisation despite not having norms.

Some tests are published with multiple norm groups, creating a bewildering choice with meaningless practical implications. The Saville Consulting Wave Focus questionnaire, which takes just 13 minutes to complete, has over 40,000 people in its norm groups, but this does not in itself guarantee validity. One could theoretically flip a coin 40,000 times as a basis for selecting people, but it is unlikely to predict their work performance effectively. Once a norm group reaches above 500 people in size, the additional insights offered are actually marginal. At this size of sample, adding further people is likely to change a sten score, a standardised scale which has a range from one to ten, by as little as 0.1 of a sten. That said, under most circumstances norms are useful in assessing people against an appropriate benchmark group, but the need for norms is very much secondary to the need for validity.

Reliability is a measure of accuracy or consistency of a test. This is usually calculated by comparing the test against itself at a different time (test-retest method); by comparing the test against another similar (parallel) version of itself (alternate form method); or by comparing some of the questions that make up the test with the other questions (internal consistency method). Ensuring high reliability is important as it improves validity, yet there remains no point in using a test that has been completed by many people and which measures each person consistently if it is completely irrelevant to their performance at work (and hence has no validity). **In essence, reliability can be thought of as "getting the test right", whereas validity is "getting the right test".**

Return on investment (or **utility**) is achieved by using a valid test in conjunction with other methods, such as a good structured interview, to select the appropriate candidates. There are different methods for calculating return on investment, but one must know the validity of the test (the correlation with job performance) and how productivity at output varies between workers. The relationship between return on investment and validity is **linear** (and not based on the square of the validity, as is sometimes reported). That is, as the validity of the measurement method goes up, so does the return on investment.

While all of these concepts are important in testing, validity is central. Possession of reliability makes a questionnaire more likely to have validity but it is not a guarantee of validity. Where a questionnaire can show, through a process of hypothesis testing, that it has superior validity this will impact on fairness and legal defensibility. Valid questionnaires lead to better decisions, fewer selection errors, more accurate identification of development needs and hence better performance of organisations and a higher return on initial investment. The first consideration for an individual deciding to use an assessment is "What is the validity, and how does this compare to the validity of other assessments?"

Because test authors tend to use very different and often ad hoc samples to demonstrate validities, it becomes virtually impossible to directly compare validity data reported from different questionnaire manuals.

Because of this and incumbent financial and resource costs, few studies have attempted to directly compare a large number of different questionnaires on the same sample, and to assess them against independent measures of performance at work. So, a study on a single sample and against the same work performance criteria was critically needed to advance knowledge in the field of personality measurement and to improve selection and development practices in the world of work.

Project Epsom

Project Epsom compared a range of the better-known personality questionnaires to determine which among them are the more valid measures of work performance. This project compared the major personality questionnaires in one study against the same job performance criteria, to create a level playing field for a direct and fair comparison.The extent to which each could measure the performance of the test-taker in a work context, as defined by both an overall measure of global performance and by the Great Eight competency framework (Kurz & Bartram, 2002), was assessed. The Great Eight framework is an independent model of work performance skill, personality, motivation and intelligence, not developed by Saville Consulting. The content of the global performance measure originates with the work of Nyfield et al. (1995) and covers three key areas: applying specialist knowledge, accomplishing objectives and demonstrating potential.

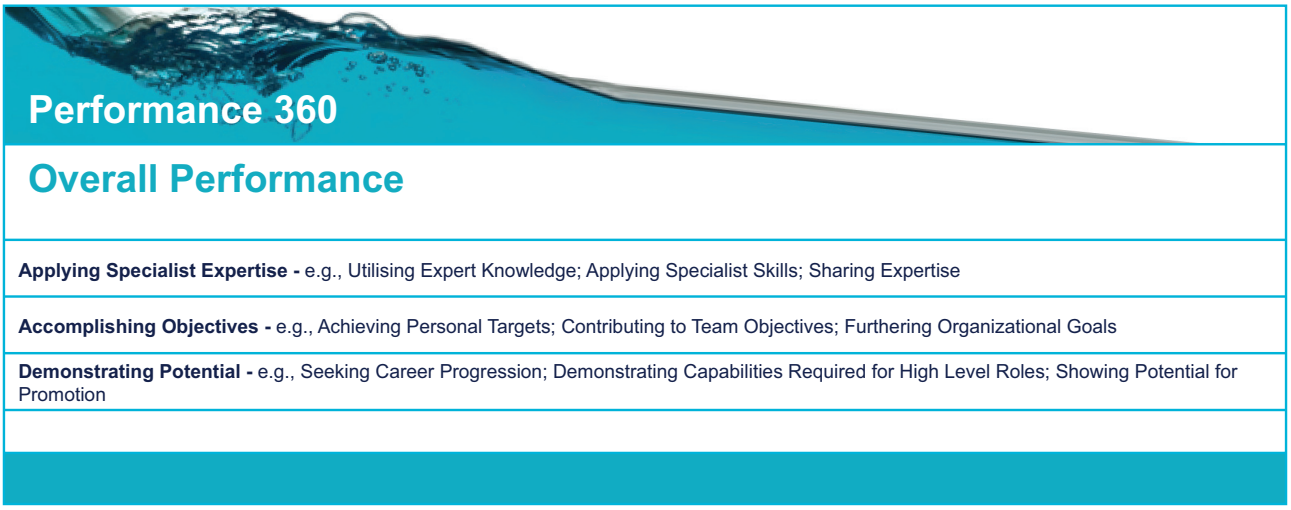
Method

A total of 308 participants completed a range of different questionnaires. In this phase one report, we consider the better-known of these, including the Professional Styles and Focus Styles versions of the Saville Consulting Wave® questionnaire, Saville PQ™, OPQ®, Hogan Personality Inventory, 16PF5 and NEO-PI-R. The majority of these participants also completed a larger range of questionnaires (29 in total), including the Hogan Development Survey, Thomas International DISC, DISCUS, and MBTI assessments. The presentation order of these questionnaires was counterbalanced across participants in order to prevent fatigue effects. Each participant was asked to nominate two other people who would act as independent "raters" and who evaluated their performance at work.

The Performance Rating Questionnaire

The Performance 360 assessment is a separate instrument from the Saville Consulting Wave questionnaires, which was designed specifically to measure work performance. It provides work performance criteria against which the different personality questionnaires used in Project Epsom can be compared. It helps to bring the field of competency measurement up to date and into the age of online business and assessment. It assesses performance completely independently of personality measurement, considering a range of different behavioural, ability and global areas of work performance. Figure 1 below illustrates the three items of global performance as presented in the Performance 360 questionnaire.

Figure 1: Measuring global performance using the Performance 360 questionnaire.



When completing the Performance 360 assessment the independent raters were asked to indicate how effective the main participant is in these and other areas on a seven-point rating scale from “Extremely Ineffective” to “Extremely Effective”. In addition to the global performance assessment, raters also provided an external rating of the performance of participants in terms of SHL’s Great Eight work competencies. Crucially, the Performance 360 assessment provided independent measures of the effectiveness of the individual in their job.

Raters were also asked to complete a personality questionnaire on themselves (Wave Focus Styles). This forms the basis of further study, looking at how the personality of the raters might influence their judgement of the work performance of others.

Initial data in Project Epsom were collected from October 2007 to February 2008 and a number of follow-up studies were run after 6 months to establish questionnaire predictive validities over time. Participants were paid for their involvement in this project and were invited to cooperate across a wide range of organizations in the UK and USA, with fewer numbers of participants from Bulgaria, Canada, Germany, France, Ireland, the Caribbean, India, South Africa, Australia and New Zealand.

Analyses

All questionnaires were compared using an identical approach against the Great Eight model and the global performance measures from the Performance 360 questionnaire. There has been some misinterpretation of the methods used in this study. **This study did not correlate the various self-report questionnaires with the Wave questionnaires**. Rather, the self-report questionnaires were correlated with independently gathered work performance ratings from participants’ managers and work colleagues using the Great Eight competency model developed by SHL (Bartram, 2005), as well as a global job performance rating. We used the Great Eight framework as this is a relatively well-known model of job competencies. These ratings of work performance were collected from managers, work colleagues, family members, partners and friends who were required to have a knowledge of the participant’s behaviours at work.

It was then possible to evaluate independently which self-report questionnaires correlated best with a third party’s ratings of job performance, in terms of overall job performance and performance of core workplace competencies. The use of an external independent model provided the fairest possible means of assessing the performance of each of the questionnaires competing in Project Epsom. We compared questionnaires against the Great Eight criteria using exactly the wording of Bartram (2005) and for the OPQ32i we used the exact Great Eight equations published by SHL, in Bartram (2005). Statistical approaches such as multiple or canonical regression, which can lead to overestimates of validity, were not used.

Prior to analysis the aspects of work performance in the Great Eight model that each questionnaire should measure was hypothesised. This was based on statistical modelling and content review. Approaches such as multiple or canonical regression, which can lead to overestimates of validity, were not used.

The Saville Personality Questionnaire

Psychometric test users sometimes become attached to a favourite test being convinced that certain scales cannot possibly be measured by other questionnaires. To challenge this orthodoxy a completely new questionnaire, the Saville Personality Questionnaire (Saville PQ™), was developed. This combines modern Wave measurement technology with the same “deductive” development approach that was employed with the OPQ® nearly 25 years ago (Saville et al., 1984). The Saville PQ was developed to demonstrate the recent advances in knowledge and to see if the same level of validity as is possessed by the OPQ® could be produced in a questionnaire that takes less than a quarter of the time (some 13 minutes) to complete.

Like the Saville Consulting Wave Professional Styles and Focus Styles questionnaires, the Saville PQ also has the added advantage of providing separate measures of people’s talents and motives in a given area, as Saville Consulting research indicates that these measures need to be clearly separated. For example, our research has revealed a distinct difference between being good at and enjoying an activity, though many questionnaires confuse the two. Questions asking about motives and talents are not identified in the OPQ® as separate measures and this can cause confusion in interpretation. For example, in the normative version of the OPQ32®, the “Forward Thinking” scale has three questions asking about whether the respondent likes to forward plan and three questions asking about whether they are good at forward planning. About 60% of the OPQ® items refer to being good at an activity and 40% refer to liking an activity. Having separate measures of motivations and talents, as in the Saville Consulting questionnaires, also helps to identify the specific development needs of individuals at work.The Saville PQ also gathers normative (free rating) and

ipsative (forced choice ranking) responses within its sub-15 minute completion time. This dynamic nipsative format, also pioneered in the Saville Consulting Wave questionnaires, helps a questionnaire counteract the natural tendency of respondents to agree with the majority of statements presented to them. A respondent can agree with as many statements as they like in the free rating normative task, even repeatedly giving the highest rating possible to many questions if they so choose, but they are then required to further clarify equally rated questions by ranking these questions in terms of how much they agree with them (ipsative ranking task).

It is noteworthy that in order to generate both a normative and ipsative measure using the OPQ® portfolio the respondent would be required to complete two much longer questionnaires, which take nearly two hours in total. The Saville PQ also avoids negative questions, as research has found that such questions were significantly less reliable than positively-phrased questions (e.g. Angleitner & Lö, 1986). The Saville PQ was used for the first time in Project Epsom.

Seven Key Questionnaires: A Summary

Figure 2 below provides a summary of seven key questionnaires that are compared in this paper.

Figure 2: A summary of seven questionnaires compared in Project Epsom.

Questionnaire	Number of Questions	Typical Completion Time
OPQ32i	416	60 mins
NEO-PI-R	240	40 mins
Wave Professional Styles	216	40 mins
Hogan Personality Inventory	206	30 mins
16PF5	185	30 mins
Wave Focus Styles	72	13 mins
Saville PQ	72	13 mins

What Level of Validity Should We Expect?

Validity, the degree of relevance a test has to work performance, is normally expressed as a value between -1 and +1. This correlation coefficient indicates the extent of the relationship between the questionnaire and job performance. A validity of zero indicates a chance measurement. This is as effective as flipping a coin to predict how an individual is likely to perform at work.

A validity of 1 would be a perfect measurement of how an individual is likely to perform at work. Of course, a perfect measurement of performance is impossible as no single assessment method can account for all of the factors that constantly impact on people's performance at work. Validities in the range of +0.8-0.9 are also unlikely in the extreme to be obtained using any single method.

Studies using huge databases of information suggest that a good personality questionnaire can be expected to show validities of about +0.3, which is a very useful degree of validity. To put this into context, ability tests may have validities around +0.5, a standard job interview is likely to have validity of around +0.2 and references or educational qualifications are likely to be as low as +0.1 (Schmidt & Hunter, 1998).

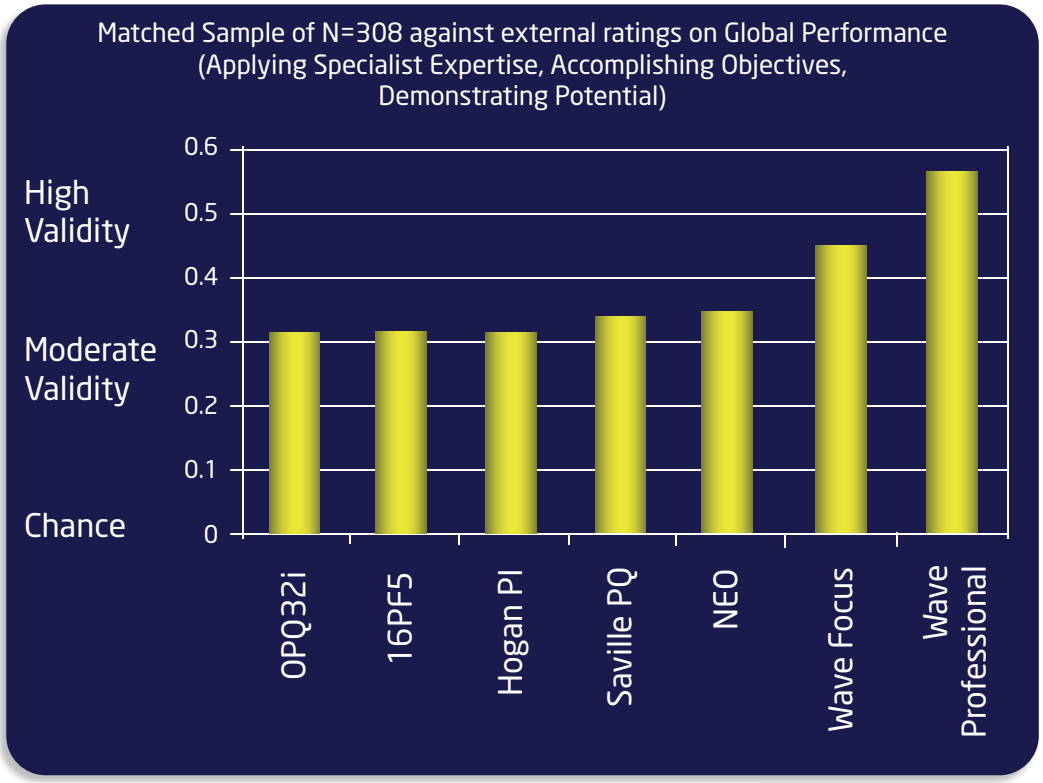
These validity figures come from a statistical procedure known as meta-analysis which takes into account such factors as the degree of unreliability inherent in obtaining various subjective ratings of job performance. Such factors had previously led to underestimates of the "true" validity of a selection method. In Project Epsom, the unreliability in the ratings of job performance obtained was statistically taken into account, but crucially we report on a complete data set where we did not exclude any data. This was done in order to ensure a standardised method across all questionnaires and to keep the playing field as even as possible.

Results Summary

Total job performance was measured through a three-item Global Performance scale (Kurz et al., 2009). Figure 3 shows the validities of seven key questionnaires in measuring global work performance, as assessed by the raters through the Performance 360 questionnaire. This global measure was chosen to ensure a standardised assessment across all of the questionnaires and represents a view of performance at work in terms of applying specialist knowledge, accomplishing objectives and demonstrating potential.

The Global Performance measure used is particularly useful as it is a general criterion which does not favor any particular personality questionnaire over the others. The more accurately we can use the responses on a given personality questionnaire to predict what an independent rater has said about the work performance of the test-taker, the more valid this personality questionnaire can be considered to be.

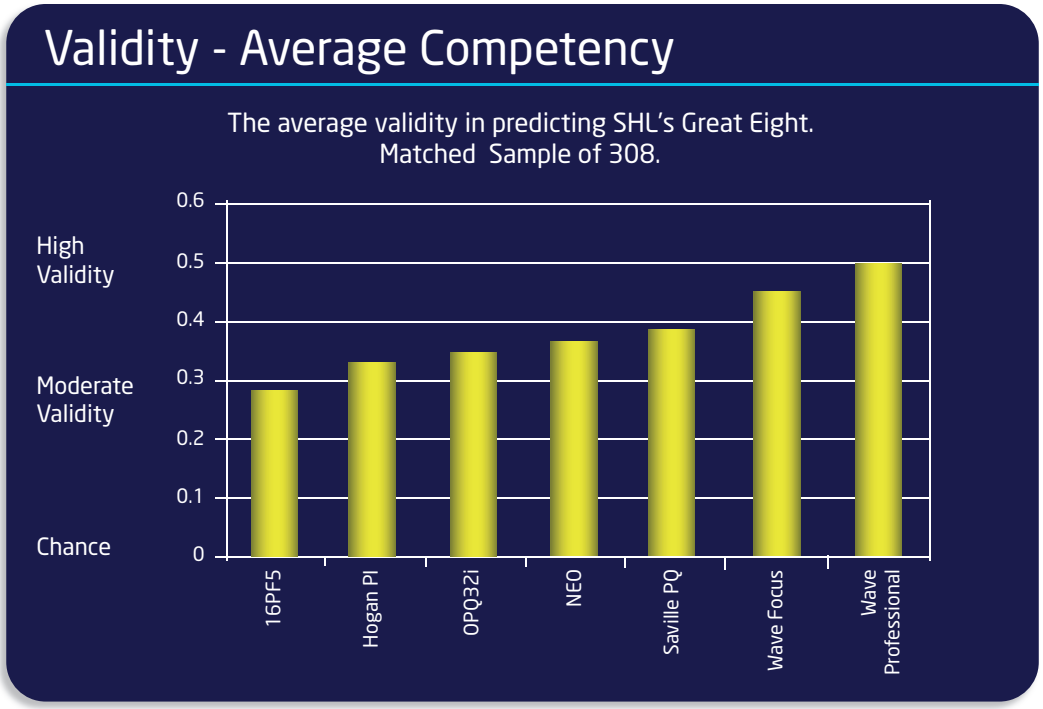
Figure 3: The validity of seven key questionnaires in measuring total job performance.



All of the seven questionnaires here showed at least a moderate level of validity in predicting work performance, considerably higher than the values considered by many studies of personality testing (e.g. Schmitt et al., 1984; Barrick & Mount, 1991; Morgeson et al., 2007). The Wave Professional Styles questionnaire eclipses all other questionnaires. The Saville PQ compares favourably to the OPQ32i despite taking just 25% of the completion time, and also is comparable in validity to the Hogan Personality Inventory and 16PF5, which take approximately twice as long.

These seven key questionnaires were also compared against external ratings of the Great Eight work performance competencies in turn. Validities were calculated for measuring each of the Great Eight competencies and these scores were then averaged together. These average validities in measuring work performance are shown below in figure 4.

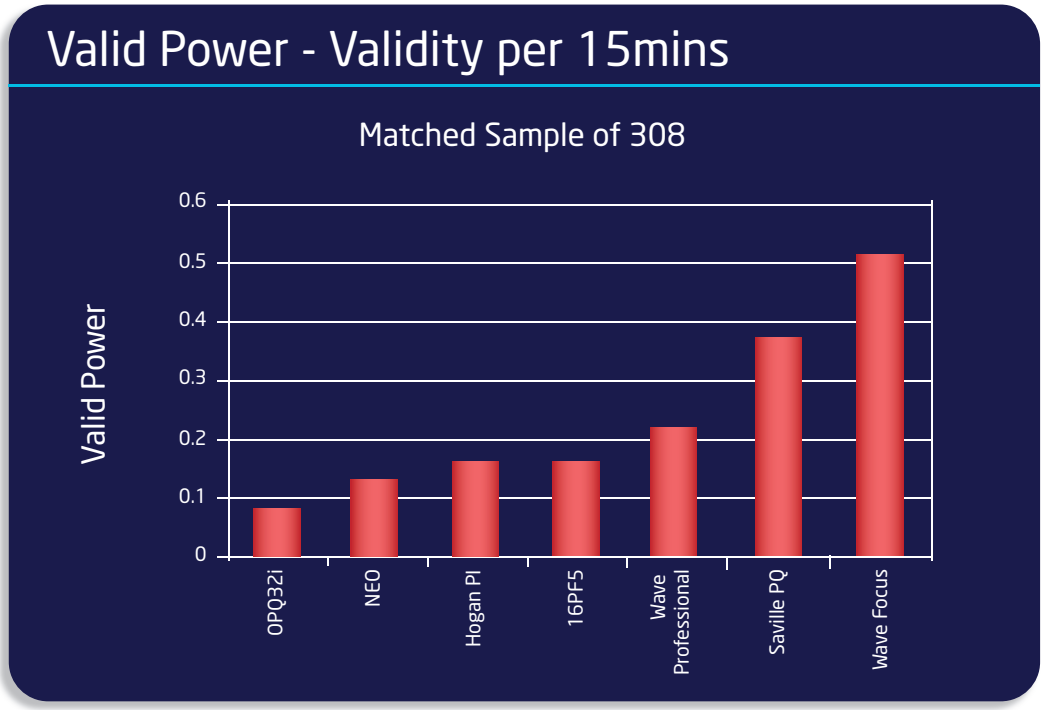
Figure 4: The average validity of seven key questionnaires in measuring the Great Eight competencies.



In terms of the Saville Consulting questionnaires, the results for the individual Great Eight competencies are thus consistent with the result for global performance.

"Power" relates to measuring effectiveness or output in a given unit of time. In terms of personality questionnaires, that which provides the greatest validity per unit of time. Figure 5 (below) compares the power of the questionnaires in terms of how much validity can be achieved by each in 15 minutes.

Figure 5: The power of seven key questionnaires in terms of their delivery of validity in 15 minutes.



As can be seen, the Wave Focus Styles and Saville PQ questionnaires are the most powerful, offering good levels of validity in the shortest completion times.

Further results from this large research project will be presented in a number of future papers to further complement the existing range of international validation studies on specific occupational groups. Saville Consulting Wave validation studies have been carried out in countries such as the UK, the USA, Mexico, Brazil, France, Denmark and Spain and have looked specifically at occupational groups including managers, engineers, consultants and civil servants. For the Saville PQ, we are pleased to report that it is showing impressive levels of test-retest and alternate form reliability over a period of six months, despite its short length. A number of other popular assessments were also included in Project Epsom and their performances are discussed in other papers. To give a flavor of these results, below we provide a brief summary of some of the key findings.

Thomas International DISC / DISCUS

The DISC model was first described in 1928 by William Marston, a "psychologist and inventor" whose greatest achievement was perhaps his creation of the cartoon character Wonder Woman. The Thomas International website stated that their version of the DISC questionnaire "measures work behaviours and is +0.75-0.95 valid". No validity was found of this magnitude, neither for Thomas International DISC nor the DISCUS variation. Indeed, no questionnaire or performance measurement method has been shown, to the authors' knowledge, to have validity at this high level.

The results for the DISC questionnaire were in line with the Buros review of DISC but curiously discrepant from the more positive British Psychological Society Psychological Testing Centre review. To quote part of the Buros review of the DISC model (Plake & Impara, 2001):

"There appears no research that finds DISC to measure the traits measured in Marston's model. The evidence does not meet the criteria established in the APA (American Psychological Association) Standards for Educational and Psychological Testing as of 1999. There are no studies that specify what the DISC predicts. The test suffers from questionable reliability and unknown validity. The use of DISC is not recommended."

Myers-Briggs Type Indicator (MBTI)

There was limited support for validity of two of the MBTI scales. There was some degree of support for responses to MBTI questions about "Extraversion" accurately measuring a person's ability to influence people at work. Similarly, there was some evidence that people who were high on "Judging" in the MBTI were seen as being better at delivering results at work.

Hogan Development Survey (HDS)

It was difficult to relate the HDS to job performance. In the HDS, if an individual scores beyond the 84th percentile in certain areas (e.g. "Sceptical") they go into the "Dark Side", an area of extreme strength identified as being potentially problematic.

When 180 participants re-completed the HDS questionnaire one week later, only 8% kept exactly the same "Dark Side" profiles. The well-established problem of having arbitrary cut-off scores in psychometric assessments is highlighted here and the situation is resonant of the 11+ assessment which was a feature of British education some years ago. Every child at the age of 11 took an IQ test, which served as a selection test to enter more or less academic senior schools. When retested two years later at the age of 13, some 35% of children fell the other side of the cut-off from where they were at 11.

Implications of Project Epsom

As has been discussed, "return on investment" is linearly related to validity. Moving from recruiting using a test with a validity of +0.2 to using a test with a validity of +0.4 can double the cost benefit because tests with a higher validity are more likely to ensure the candidates who are going to perform better are selected. Small increases in validity can also have a large effect. For example, going from using a test with a validity of +0.3 to +0.4 is a 25% increase, which can have a huge impact on the organisation's productivity and hence on return on investment.

A key organisational requirement may be to identify your top 20% of performers (those with high potential). Figure 6 below illustrates the decrease in selection errors which will be incurred as validity is increased. The figure may vary of course from situation to situation, but this offers a general guideline.

Figure 6: The effect of validity on probability of selection errors.

Identifying High Fliers

	Validity	By selecting based on top 20% of questionnaire completers...
No Validity	0	**4** out of the 20 will prove to be in the top 20% of job performers
Moderate Validity	.3	**7** out of the 20 will prove to be in the top 20% of job performers
High Validity	.6	**10** out of the 20 will prove to be in the top 20% of job performers*

**17 out of 20 will prove to be above average performers*

	Validity	By selecting based on top 20% of questionnaire completers...
No Validity	0	1 person in **5** selected will prove to be bottom 20%
Moderate Validity	.3	1 person in **10** selected will prove to be bottom 20%
High Validity	.6	1 person in **50** selected will prove to be bottom 20%

Particularly serious errors occur when someone from the bottom 20% is identified as demonstrating top 20% potential. Once the validity of your assessment method is as high as +0.6, only one in every fifty of the bottom 20% performers will be incorrectly selected as demonstrating top 20% potential. In other words, tools with a higher validity dramatically reduce the frequency of serious selection errors. This simple example shows the importance of using tests which are valid and suitable for predicting actual performance at work, such as the Saville Consulting Wave Styles questionnaires. These were designed specifically to maximise prediction of performance and potential at work. It is not sufficient for test publishers merely to cite the degree of agreement between their test and another as the validity of the test. Tests must be designed to actually relate to real-life performance measures: with the study reported here, this is job performance.

Why are the Saville Consulting Wave questionnaires more valid?

Up until very recently, the academic consensus was that the highest obtained validities in measuring job performance for personality tests were low, relative to those obtained by other tools (e.g. Hurtz & Donovan, 2000). This belief pervades the field of psychometrics and academics implore test publishers to "provide a theoretical validation of their measures" (Ferguson, Payne & Anderson, 1994).

The recently-developed Saville Consulting Wave Styles range (see MacIver et al., 2006a) is a modern job-relevant measure of personality which provides just such a theoretical validation of its measures and obtains validities in measuring job performance in excess of other popular questionnaires. There are two Wave questionnaires: the Wave Professional Styles questionnaire which is most commonly completed by respondents in about 40 minutes to complete and the 13 minute Wave Focus Styles questionnaire, which covers the most valid questions from the Professional Styles questionnaire and thus maintains about 80% of the validity of the fuller assessment. Both are built on the same Wave Styles framework. Below is a summary of key reasons why the Saville Consulting Wave Styles questionnaires are revealed to be the most valid personality questionnaires.

1. The Wave questionnaires are modern and are written in the language of contemporary business

In taking into account recent advances in computerisation and the internet, they measure personal characteristics and competencies which are relevant for business today. For example, the Wave questionnaires measure inclination to use information technology, which has become a huge part of many

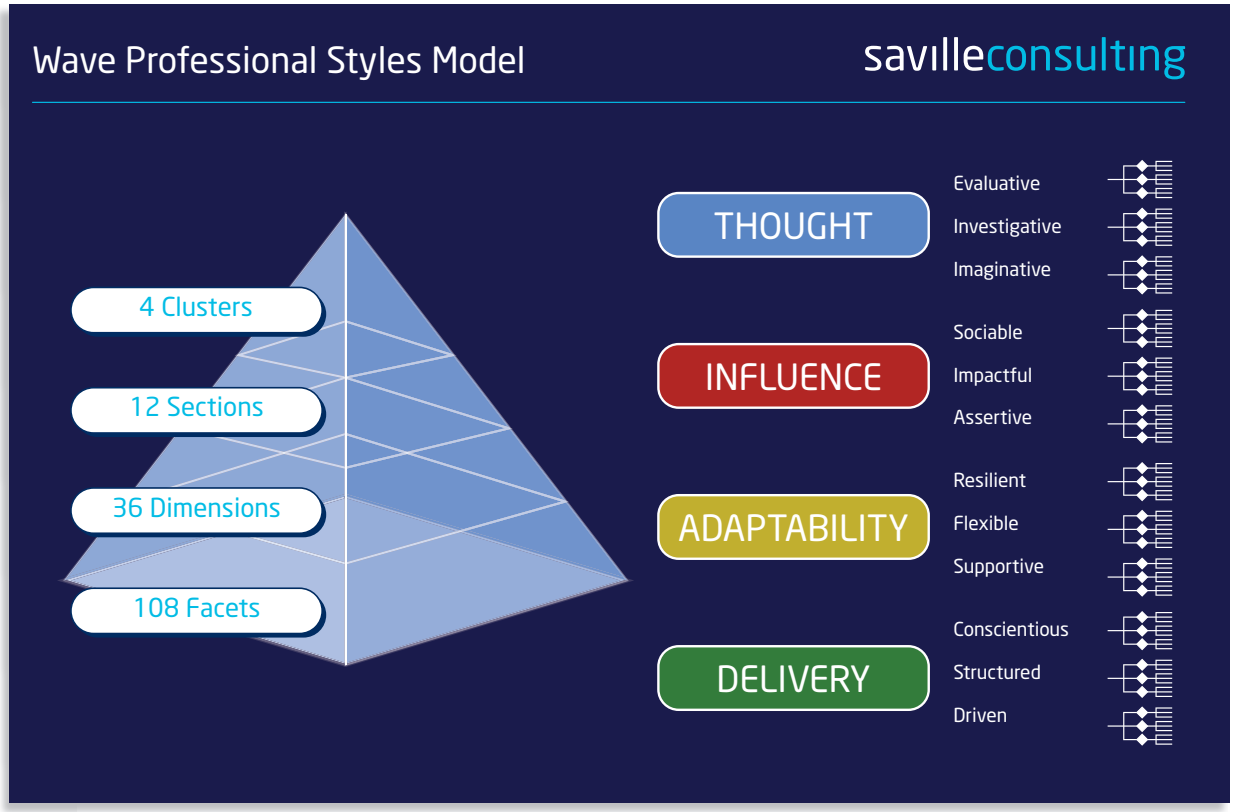
people's daily work. Questionnaires developed in previous decades, naturally, have not taken into account the changing nature of job roles and work culture and so may be measuring out of date aspects of work performance. A radical rethinking of the measurement base was warranted. Many other questionnaires are not related to industry at all, so it is little surprise that they are shown to be less valid than work-relevant assessment methods in measuring work performance (Robertson & Smith, 2001).

2. *The Wave questionnaires were developed through an extensive understanding of the field of personality assessment and work performance*

The Wave questionnaires were continuously validated during their construction, combining different development strategies (Saville et al., 2009b; Saville et al., 2009c). Individual items were validated against work performance criteria from conception and external ratings of work performance were obtained, so it was known which questions would measure which competency from the outset. Older questionnaires tend to focus on validity only at the more general scale level, whereas validity was written into the individual items of the Wave questionnaires from the very beginning (see MacIver et al., 2006b). This ensured that when the individual questions were combined to form scales, the scale validities were greatly enhanced too.

Professor Peter Saville, author of the OPQ®, built on his extensive knowledge of the domain and combined this with online trialling of the questionnaires to select the most valid questions while his development team produced an overarching Wave framework considering personality, competencies, aptitudes and the impact of culture in the workplace. Kurz et al. (2008) showed how the Big Five personality factors and Great Eight competencies align with the Saville Consulting four Behavioural Performance Clusters and twelve Behavioural Performance Sections in the Wave model. Figure 7 below provides a graphical summary of the hierarchical nature of the Wave Professional Styles framework, moving from four overall clusters down to 108 facets. Each one of these facets is measured through a motive and talent question, meaning that the questionnaire has a total of 216 items.

Figure 7: The Saville Consulting Wave Professional Styles model.



Extensive research on personality questionnaires has indicated that certain strategies are key to successful questionnaire design and these were capitalised upon in the development of the Wave framework. For example, it is known that many questionnaires are full of badly-worded, ambiguous and/or negatively-phrased questions (Angleitner & Lö, 1986) which can reduce a questionnaire's validity. It is often the smallest words that can cause the biggest problems. "Ifs", "buts" and "ands" complicate questions by allowing ambiguities in interpretation. Similarly, shortening a measurement scale in a questionnaire to just a few questions can actually improve its validity (Burisch, 1997).

Hitting the core of a concept with good items is more productive in questionnaire construction. Three well-written and direct items can achieve the same (if not a greater) level of validity when compared to a large number of poorer items (Burisch, 1997). This “keep it simple” approach was demonstrated recently by Lie (2008) who showed that one item can screen for excessive daytime sleepiness as effectively as a full day of physiological and psychological tests. That one item was: “Measure your sleepiness on a typical day, where 0 = none and 10 = highest”. Looking at the completion experience from the respondent’s perspective, it is no surprise that people can become bored and inconsistent in their responses if they are asked the same question a dozen times. The Wave questionnaires are built around extensively validated items from concise scales which provide clear, discerning links to work performance (Saville et al., 2009b; Saville et al., 2009c).

3. *The Wave model introduces a number of groundbreaking features which provide breadth of information and sophisticated distortion detection*

The dynamic rate-rank format combines ipsative and normative assessment in one interative online measure and helps tohelps identify how consistently people have answered the questionnaire. This format also counteracts the natural tendency of people to agree with most statements presented to them. It therefore offers a more judicious means to discriminate people’s preferences and styles of behaviour (Saville et al., 2009b; Saville et al., 2009c). It is also a useful tool for assisting in identifying situations where respondents are attempting to manage the impression they are making in the questionnaire. For example, where the ipsative score is very much lower than the normative score, our experience indicates that the candidate may have exaggerated in the free rating (normative) task. This provides a much more sophisticated measure than an overall social desirability scale as we can pinpoint specific areas to discuss in the report, rather than being left unsure where or why people are responding in a socially desirable way.

There is the problem with normative-only questionnaires that many people use the middle of the scale, which others continually use the extremes. This is known as central tendency response set. This makes the comparison between people, for example in different nationality groups, problematic. As quoted by Saville and Wilson (1991), Simpson, for example, found that the word “frequently” meant “over 80% of the time” to some participants and “under 40% of the time” to others. It is also not unknown for some people to use extremes on a 5-point scale over 50% of time, while others never do. Ipsative questionnaires force a degree of negative correlations between scales, but it has been shown that with more than about sixteen scales this effect is minimal (Baron, 1996). There is also some evidence that ipsative scales can help to control distortion better than the normative format. Both normative and ipsative scales have inherent response biases and it is to counteract these limitations that the two methods were combined in the Saville Consulting questionnaires, improving validities by some 10%.

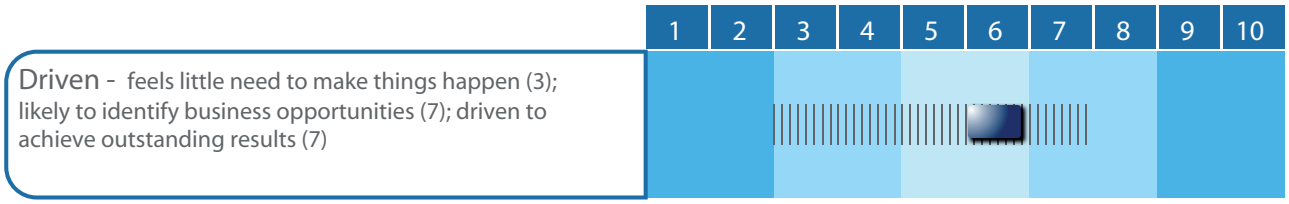
The Wave questionnaires (and Saville PQ) also exist in parallel forms, so there are two alternate and reliable versions of the questionnaire available for use. Thus, if there is some concern that a candidate is distorting their responses when completing an unsupervised questionnaire online, they can later complete a parallel version of the same questionnaire in a supervised format, affording comparison of their responses across the two completions. Nevertheless, it is interesting (and encouraging) that several meta-analyses have shown that candidate distortion in personality assessment is far less prevalent and affects the validity of responses less than has previously been assumed (e.g. Hough et al., 1990; Ones et al. 1996; Ellingson & Sackett, 2001; Schmitt & Oswald, 2006).

Separate motive and talent measures in the Wave questionnaires help identify specific areas where people are motivated to improve (where their responses to questions about their motivations are discernibly higher than their responses to questions about their talents in the same area). Similarly, they allow us one to see areas where a person may be less motivated, even if their talent (and hence outward behaviour) is not demonstrably low (where their responses to questions about their talents are discernibly higher than their responses to questions about their motivation in the same area). Figure 8 below illustrates motive-talent splits in a Wave report. In this case, the M (motive) marker is somewhat higher than the T (talent) marker, indicating the respondent is more motivated to act in an enterprising manner than they feel they currently are showing. This might well be an area that their manager could focus on developing with them, for example by identifying which elements of their work environment may be impacting on their performance.

Figure 8: A motive-talent split.

The Wave questionnaires show facet ranges where the respondent has answered closely-related questions somewhat differently. This aspect of the response style is shown in the report as hatching marks. Below, in Figure 9, is an example of a facet range from the Wave Focus Styles profile of Ian Woosnam OBE, the captain of the European Ryder Cup 2006 team and formerly the number one ranked golfer in the world.

Figure 9: A facet range.



While Ian is driven to achieve results and is likely to identify the means to do this, it is interesting to see that he reports less need to make things happen. This has resulted in a facet range in the Wave profile, where Ian has rated the questions that relate to business opportunities and achievement of results significantly higher than he has rated the questions about making things happen. Consequently, the facet range is a means to highlight a unique aspect of a given individual's preferred behavioural style. It is possible that for Ian, his success throughout his career has meant that he's never explicitly felt a need to consciously make things happen; instead, his inherent talents have meant that things simply do happen for him. More generally, facet ranges add further richness of information and guide discussion about the unique responses of an individual.

4. The Wave questionnaires make their validities readily available to users and are easy to understand

The Wave questionnaires are administered and scored online to maximise ease of use for both candidate and practitioner. They also make their improved validities easily available to users. Being able to examine the validity of a questionnaire with statistical techniques does not necessarily make the validity meaningful to the user. Valid scores must be presented simply in a report or profile which can be easily interpreted and used as a direct measure of work performance (MacIver et al., 2008). Through the sophisticated yet user-friendly Wave reports, practitioners are provided with an even better understanding of people's behaviours and performance at work than has previously been available to test-users and even Industrial Psychologists.

Discussion

Project Epsom is one of the first studies to compare the criterion-related validities of a range of popular personality questionnaires in a single database. It also confirms existing research carried out on personality questionnaires. For example, the validities demonstrated here for the Hogan Personality Inventory are consistent with the meta-analysis by Hogan et al. (2008) and a study by Foster et al. (2008). The OPQ32i validities for the individual Great Eight competencies are similar to those reported by Bartram (2005), but the Great Eight total validities are considerably lower. Possible reasons for this are explored below.

In Project Epsom, we report on a full data set and did not employ multiple regression or somewhat dated statistical techniques such as canonical analysis (as used by Bartram, 2005), which can produce serious over-estimates of validity. Canonical analysis can be thought of as juggling both the predictor (e.g. personality questionnaire scales) and the work performance criteria until some optimum equation is found which maximises the correlation between the predictor and criteria. The danger is that canonical analysis could lead us into a situation rather like the theory suggesting that an infinite number of monkeys using

typewriters would ultimately produce the complete works of Shakespeare by chance. That is, the procedure keeps juggling the data until a "best-fit" solution is reached by chance. When conducted on ipsative (forced ranking) personality scales, canonical analysis, with its enormous tendency to capitalise on chance effects, is a particularly deadly cocktail. No research is known to the authors where unreplicated canonical analysis on ipsative personality scales (as in Bartram, 2005) has produced a statistically legitimate outcome.

For example, based on canonical analysis, SHL has suggested that "the true overall combined validities of the OPQ32 (ipsative version)…actually achieves +0.55". What has possibly happened here, with respect, is the classic "Popcorn Effect", where the method has selected the popcorn thrown up highest by the popcorn machine, on a chance basis. Without replication in further samples, the serious danger of "Blunderbuss-" or "Shotgun-Empiricism" exists, where the target is indiscriminately blasted until any hit is registered.

The deep seated problem of using canonical analysis and related multivariate techniques with ipsative data may help explain the counter-intuitive formula by Bartram (2005) which **negatively** weighted emotional stability in order to measure job success. This completely conflicts with the worldwide research literature which shows that it is the emotionally stable people who tend to be more successful in jobs, and not the seriously neurotic and disagreeable. In the words of Cronbach (1970): "The investigator should be particularly skeptical of weights that make little psychological sense, since they are likely to have come from sampling errors". In this case, it is possible that the ipsative nature of Bartram's data, combined with the way multivariate statistical techniques capitalise on chance, has led to this counterintuitive finding. Cronbach is quick to point out that "simply assigning equal weight to all relevant tests… often works just as well in the next sample". Without replication in a new (cross validation) sample, it is very difficult to ensure that such a counterintuitive finding is a valid one. In Project Epsom, as has been found in many other studies, results revealed that emotional stability was positively related to overall work success.

This is not to say that ipsative questionnaires with sufficient scales cannot be treated statistically. However, much more care and caution is advised (Saville & Wilson, 1991), with replication in separate samples. Canonical analysis and multiple regression are post hoc (after the event) techniques, where the known results are further investigated in order to maximise validity. Predicting what stocks and shares were worth yesterday is rather easier than predicting what they will be worth tomorrow! Improperly used, canonical analysis can therefore be the ultimate "fishing trip" where results are simply cherry-picked as suited. Canonical analysis is now considered by many to be a rather outmoded practice, superseded by superior methods. A more discerning practice is to designate a priori (before the event) which questionnaire scales should correlate with which measures of job performance and then test these specific hypotheses. In Project Epsom, this is precisely what was done.

In personality research, as with many areas of scientific enquiry such as medical research, there is also the phenomenon of the **"file-drawer effect"** (e.g. Allison et al., 1997; Bauchau, 1997; Scargle, 2000). This effect describes the publication bias where only positive results are reported and published. If, for example, one were to develop a structured interview which demonstrates validity in measuring how successfully candidates are expected to perform in their job, these results are likely to be published as an academic paper. Where the interview is shown to demonstrate no validity, it is less likely to be written up and may be filed away. Similarly, where no correlations are found between a questionnaire and job performance the results are may simply not be published. Goldacre (2008) relates in his book, Bad Science, a quotation from Francis Bacon who noted that "it is the peculiar and perpetual error of the human understanding to be more moved and excited by affirmatives than negatives".

In meta-analysis, where many existing studies are amalgamated into a large database for further study, if only affirmative results are included this could lead to overestimations of the validity of a given procedure. With Bartram's (2005) meta-analytical data, this file-drawer problem may be one of a number of methodological issues, because the majority of the studies considered by Bartram (2005) are likely to have been conducted solely by SHL companies. It is not clear whether all studies considered were included in his final results. In Project Epsom, the complete results from one database are reported. Meta-analysis itself is an extremely powerful statistical procedure and in the opinion of the authors, the researchers who developed this methodology, Hunter and Schmidt, should have been considered for a Nobel Prize. Meta-analysis has had an enormous impact not only in psychology but also in many other areas of scientific research, including medicine. It is nevertheless important that all adequate studies are included in meta-analysis, not merely those which have been selected to support one's initial hypotheses.

The Great Eight model is one independent framework against which to measure the validity of all questionnaires on a level playing field. However, it is not necessarily the ultimate method. Against the more discerning Twelve Behavioural Performance Sections in the Saville Consulting Wave Performance Culture Framework, the Wave Styles questionnaires actually outperformed all other questionnaires by about 50%.

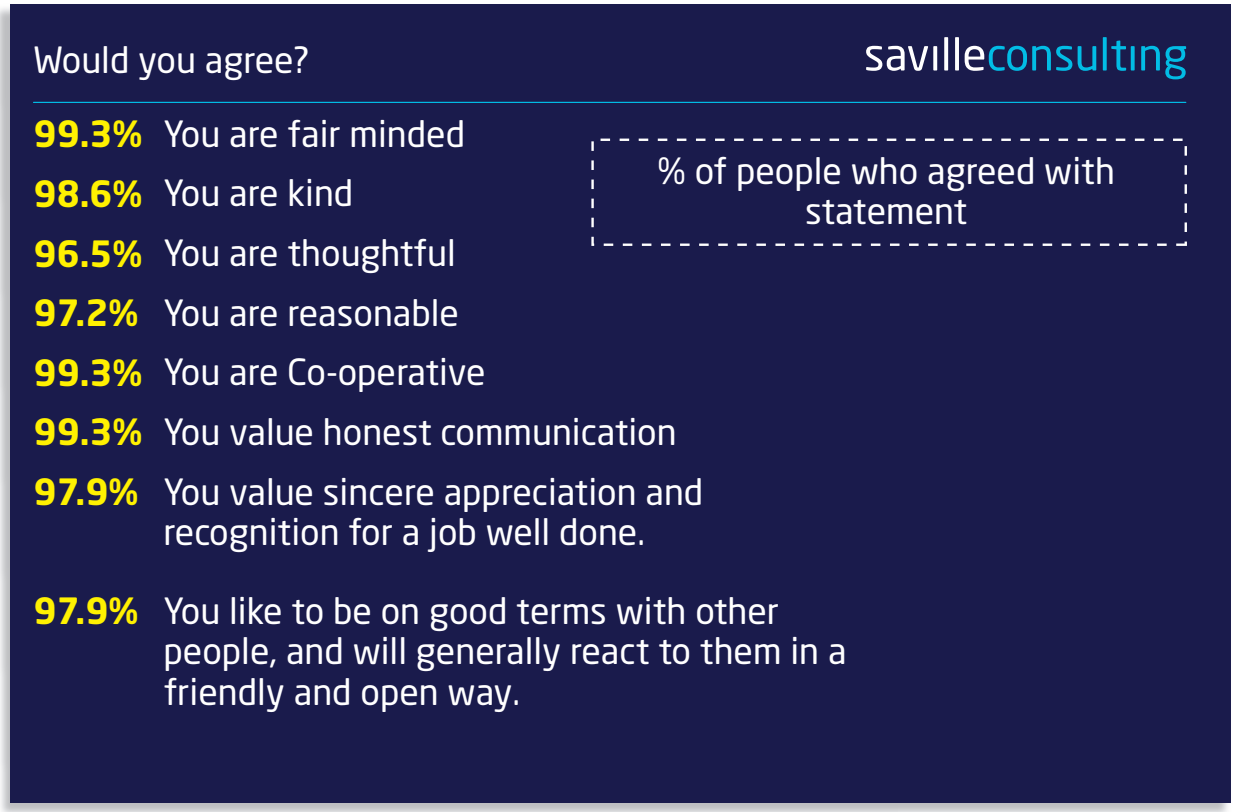
This model was not used here in order to ensure absolute fairness across all questionnaires, but this model of work effectiveness is the subject of ongoing research.

Project Epsom clearly indicates that the Saville Consulting Wave Professional Styles predictions of competency demonstrate the greatest validities in measuring global job performance. It is rewarding to see such newly developed measures outperforming traditional measures by showing considerable increases in validity. The Saville Consulting Wave Styles questionnaires make use of contemporary technological and scientific advances in measurement to establish their position as the most effective personality questionnaires for measuring job performance.

The Saville PQ™ occupies the middle ground between the Saville Consulting Wave® questionnaires and other prominent personality assessments in the market. The Saville PQ is available online with no licence fee to pay, and offers users familiar with the OPQ® the chance to use a comparable questionnaire which is more valid against the SHL criteria, takes less than 15 minutes, gives both normative and ipsative scores in that time and provides separate measures of motives and talents. In terms of reliability of measure when compared to the OPQ32i, the combined normative and ipsative scales of the Saville PQ when mapped to their like-for-like OPQ32i counterparts have an average correlation exceeding +0.7. Nevertheless, as with the OPQ®, the Saville PQ does rely on 1980s development methods and even better decision making tools are offered in the Wave Focus or Wave Professional Styles questionnaires.

However, far from selecting the most valid measures of work performance, it seems that users of some questionnaires become attached to the tests that "look right" or appear to be appropriate, which is known as faith validity (Saville, 1975) This aspect of validity is where a test user becomes familiar and happy with a tool and is very resistant to change, even though there may be more empirically valid modern alternatives available. To demonstrate this tendency we returned to a classic experiment by Stagner (1958) which documents the "Barnum Effect". A series of statements were taken or adapted from the **actual feedback reports** of certain of the personality questionnaires used in Project Epsom and 144 participants were asked whether they thought the statements were an accurate description of them. Figure 10 below indicates the percentage of people who thought each statement was accurate for them.

Figure 10: Percentage of people who agreed that each statement accurately described them.



Almost everyone said that these statements provided an accurate description of them. If the end result of some personality questionnaires is such generalised and socially desirable reports, it is perhaps no wonder that there is a perplexing level of appeal to these questionnaires which seems at odds with how poorly they differentiate people, and the limited validity they demonstrate in measuring performance at work. It seems

that some questionnaire publishers confuse the number of people who think that a report is accurate with the validity of their questionnaire, but possession of faith validity does not mean that a questionnaire measures job performance well. As Descartes once remarked:

"Common sense is the best distributed commodity in the world, for every man is convinced that he is well supplied with it"

Saville (2008) notes that where there is an unusual result in a study or questionnaire manual it is often explained away by psychologists with "neologistic gobbledigook", when in fact it is a simple error of scoring or in the data analysis. He refers to the effect known in the philosophy of science as the "Crabtree Bludgeon", where "no set of mutually inconsistent observations can exist for which some human intellect cannot conceive a coherent explanation, however complicated".

Personality questionnaires are actually viewed as useful and acceptable methods for selection by the public at large. In a sample of one thousand participants, we found that only 13% of people felt that personality questionnaires were not effective for selection, whilst 26% of people were critical of the interview. The acceptability of various techniques in a selection context is shown below in figure 11.

Figure 11: The acceptability of different selection techniques.

Acceptability of Selection Techniques (N=1000)

savilleconsulting

Useful?	"Yes"	"Uncertain"	"No"
Interviews	41%	34%	26%
References	49%	37%	14%
Intelligence	56%	27%	17%
Personality	53%	34%	13%
Astrology	22%	28%	50%

While the results indicate that personality questionnaires on the whole are acceptable to people and their use is surprisingly welcomed, this does not in itself indicate that personality questionnaires measure work performance well. It is reassuring also to see that evidence in the literature suggests that faking in questionnaires does not materially affect their validities (e.g. Hough et al., 1990; Ones et al., 1996; Ellingson & Sackett, 2001) and may actually have "minimal effects" (Schmitt & Oswald, 2006). Ellingson & Sackett describe their results as providing "additional evidence in support of a growing literature that the incidence of applicant faking is lower than might be assumed".

While some users develop a rigid loyalty to one particular test, as practitioners making significant decisions about the careers and well-being of others it is important to consider the specific functions for which any test is used. Some, like the Saville Consulting Wave Styles questionnaires, have been designed to maximise the prediction of performance at work, while others were not originally designed to be used in industrial settings at all. Some have clinical origins and bring obscure, irrelevant and badly-written items to the measurement of performance in the workplace.

By developing and sharing an understanding of performance and potential at work, it is possible to help more people self-actualise at work and to ensure that personality assessment will continue to become more efficient and valid in the future. Unfortunately, however, people continue to make extrapolations and statements from personality questionnaires which are backed up by no data whatsoever. Especially in a world where feedback and transparency of information is increasingly sought, we would do well to follow the guidance Shakespeare offers in his play Othello, The Moor of Venice:

"Speak of me as I am; nothing extenuate,
Nor set down aught in malice"

More recent predictive research on a subsample of 108 participants has also been carried out and the validity rank order of the seven key questionnaires considered in this paper was essentially maintained over a period of six months.

Of course, personality questionnaires are only one of the tools available for use. While Project Epsom shows that personality questionnaires which are modern and well-written can be valid measures of work performance, they should be used in conjunction with other techniques such as structured interviews, ability tests, in-tray tasks and job sample exercises. Indeed, personality questionnaires can be a most useful basis for structuring an interview through feedback, as part of a multi-method approach. One criticism levelled at personality assessment is that it is "just a self-perception" that has no practical application. The demonstration that personality questionnaires do have validity in measuring performance at work does not support this criticism. Aldous Huxley sagely and elegantly sums up the value of measuring personality in his 1954 work The Doors of Perception:

"To see ourselves as others see us is a most salutary gift. Hardly less
important is the capacity to see others as they see themselves"

The authors would like to thank all the participants who took part in this study and look forward to future independent replication of this work in new and separate samples.

This Project Epsom phase one report considers the results of some of the key personality questionnaires studied in this project. As this is such a large data set, further results from Project Epsom are to be released in peer reviewed academic presentations and papers over the coming year. For example, research is underway investigating the validities of further questionnaires and how the personality of raters may affect the ratings they give.

If you have any questions or comments, or would like further details about this research, please call +44(0)1372 475700 or +44(0)1534 726820. Further information and contacts can also be found at www.savilleconsulting.com/howvalidisyourquestionnaire

References

Allison, D. B., Faith, M. S., & Gorman, B. S. (1996). Publication bias in obesity treatment trials? International Journal of Obesity, 20, 931-937.

Angleitner, A., J., O. P., & Lö, F. J. (1986). It's what you ask and how you ask it: An itemmetric analysis of personality questionnaires. In A. Angleitner and J. S. Wiggins (Eds.), Personality assessment via questionnaires (pp. 61–107). Berlin, Germany: Springer-Verlag.

Baron, H. (1996). Strengths and limitations of ipsative measurement. Journal of Occupational and Organizational Psychology, 69, 49-56.

Barrick, M.R. & Mount, M.K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. Personnel Psychology, 44, 1-25.

Bartram, D. (2005). The Great Eight Competencies: A criterion-centric approach to validation. Journal of Applied Psychology, 90, 1185-1203.

Bauchau, V. (1997). Is there a "file drawer problem" in biological research? OIKOS, 19, 407–409.

Burisch, M. (1997). Test length and validity revisited. European Journal of Personality, 11, 303-315.

Cattell, R. B. (1965). The Scientific Analysis of Personality. London: Penguin.

Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. Psychological Assessment, 4, 5-13.

Cronbach, L. J. (1970). The Essentials of Psychological Testing (3rd Ed.). New York: Harper & Row.

Digman, J. M. (1997). Higher-order factors of the big five. Journal of Personality and Social Psychology, 73, 1246-1256.

Ellingson, J. E., & Sakett, P. R. (2001).Consistency of personality scale scores across selection and development contexts. Poster Session, Society for Industrial and Occupational Psychology, San Diego, California. USA.

Ferguson, E., Payne, T., & Anderson, N. (1994). Occupational personality assessment: Theory, structure and psychometrics of the OPQ FMX5- student. Personality and Individual Differences, 17(2), 217-225.

Foster, J., Johnson, C. & Gaddis, B (2008). The predictive validity of personality: New methods produce new results. Presented at the 23rd annual conference of the Society for Industrial-Organizational Psychology, April 2008.

Goldacre, B. (2008). Bad Science. Fourth Estate: London.

Hogan, J., Davies, S. & Hogan, R. (2008). Generalizing Personality-Based Validity Evidence. In S. Morton McPhail (Ed.), Alternative Validation Strategies: Developing New and Leveraging Existing Validity Evidence (181-233). Jossey Bass: San Francisco, California.

Hough, L., Eaton, N. K., Dunnette, M. K., Kemp, J. D., & McCloy R. A. (1990). Criterion related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology, 75(5), 581-595.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The big five revisited. Journal of Applied Psychology, 85, 869-879.

Huxley, A. (1954). The Doors of Perception. New York: Harper & Row.

Kurz, R. & Bartram, D. (2002). Competency and individual performance: Modelling the world of work. In I. T. Robertson, M. Callinen and D. Bartram (Eds.), Organizational Effectiveness: The Role of Psychology. Chichester: Wiley.

Kurz, R. MacIver R. & Saville, P. (2008). Coaching with Saville Consulting Wave™. In: Passmore. J. (Ed): Psychometrics in Coaching. Kogan Page, London.

Kurz, R., Saville, P., MacIver, R., Mitchener, A., Parry, G., Oxley, H., Small, C., Herridge, K. & Hopton, T. (2009). The structure of work effectiveness as measured through the Saville Consulting Wave ® Performance 360 'B-A-G' Model of Behaviour, Ability and Global Performance. (In press).

Lie, D. (2008). A Single Subjective Question May Help Screen for Excessive Daytime Sleepiness. Journal of Clinical Sleep Medicine, 4,143-148.

MacIver, R., Saville, P., Kurz, R., Mitchener, A., Mariscal, K., Parry, G., Becker, S., Saville, W., O'Connor, K., Patterson R., Oxley, H. (2006a). Making Waves: Saville Consulting Wave Styles questionnaires. Selection and Development Review, 22(2), 17-23.

MacIver, R., Saville, P., Kurz, R., Henley, S., Mitchener, A., Mariscal, K., Parry, G., Becker, S., Hurst, E., Saville, W., O'Connor, K, Patterson R., McLellan, S. & Blakesley, M. (2006b). The Validation Centric Development of the Professional Styles Questionnaires. Presented at the BPS Occupational Psychology Conference, Glasgow, UK.

MacIver, R., Saville, P., Kurz, R., Anderson, N. & Evers, A. (2008). New Aspects of Validity: User Available and User Received Validity. Presented at the ITC Conference.

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(3), 683-729.

Nyfield, G., Gibbons, P. J., Baron H, & Robertson, I. (1995). The Cross Cultural Validity of Management Assessment Methods. Paper presented at the 10th Annual SIOP Conference Orlando, USA, May 1995.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). The role of social desirability in personality testing for personnel selection: The Red Herring. Journal of Applied Psychology, 81, 660-679.

Plake, B. S. & Impara, J. C. (Eds.) (2001). The Mental Measurements Yearbook, University of Nebraska Press, Lincoln: USA.

Robertson, I. T. & Smith, M. (2001). Personnel selection. Journal of Occupational and Organizational Psychology, 74(4), 441-472.

Saville, P. (1975). Occupational Testing. HDS Management Consultants, London.

Saville, P. (2008a). Personality Questionnaires – Valid Inferences, False Prophecies. Presented at the Division of Occupational Psychology of the British Psychological Society Annual Conference, UK, January 2008.

Saville, P. (2008b). Does Your Test Work? Presented at the Psychological Society of South Africa Annual Conference, Johannesburg, August 2008.

Saville, P. (2008c). A Comparison of Leadership in Business and Elite Athletes. Presented at the A&DC Conference, Institute of Directors, London, November 2008.

Saville, P., Holdsworth, R., Nyfield, G., Cramp, L., & Mabey, W. (1984). Occupational Personality Questionnaire Manual. Thames Ditton: Saville-Holdsworth, Ltd.

Saville, P, MacIver, R., Kurz, R. & Hopton, T. (2009a). Saville PQ Manual and User Guide. Saville Consulting Group: Jersey. (In press).

Saville, P., MacIver, R. and Kurz, R. (2009b). Saville Consulting Wave Professional Styles Manual and User Guide. Saville Consulting Group: Jersey. (In press).

Saville, P., MacIver, R. and Kurz, R. (2009c). Saville Consulting Wave Focus Styles Manual and User Guide. Saville Consulting Group: Jersey. (In press).

Saville, P. & Wilson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of personality. Journal of Occupational Psychology, 64, 219-238.

Scargle, J. D. (2000). Publication bias: The “file-drawer” problem in scientific inference. Journal of Scientific Exploration, 14(1), 91-106.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research. Psychological Bulletin, 124(2), 262-274.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Meta-analyses of validity studies. Journal of Applied Psychology, 70, 280-289.

Schmitt, N. & Oswald, F. L. (2006). The impact of correction for faking on the validity of noncognitive measures in selection settings. Journal of Applied Psychology, 91(3), 613-621.

SHL (1999). OPQ32 Manual and User’s Guide. Surrey, UK. SHL Group plc.

Stagner, R. (1958). The gullibility of personnel managers. Personnel Psychology, 11, 347-352.

Wiggins, J. S. (1973). Personality and prediction: Principles of personality assessment. Reading, MA: Addison-Wesley.

www.savilleconsulting.com/howvalidisyourquestionnaire

For more information please call

+44 (0)1372 475700

Saville Consulting UK Ltd
Harley House
94 Hare Lane
Claygate
Surrey
KT10 0RB
United Kingdom
Tel: +44 (0)1372 475700
info.uk@savilleconsulting.com

Saville Consulting British Channel Islands
1st Floor
Anley House, Anley Street
St Helier
Jersey
JE2 3QE
Tel: +44(0)1534 726820
info.group@savilleconsulting.com

Saville Consulting Asia Pacific
P O Box 1855
North Sydney
NSW 2060
Australia
Tel: +612 8004 2941
info.ap@savilleconsulting.com

Saville Consulting North America
P.O. Box 1446
Hightstown, NJ 08520
USA
Tel: +1 609 918 9009
info.us@savilleconsulting.com

Frequently Asked Questions

Since the first version of this paper was released, there has been great international interest in this research. Some of the more commonly asked questions are addressed below.

Why was the research conducted?
To improve knowledge and validity in the assessment field. It is extremely difficult to meaningfully compare the validity of different assessments on very different groups of people. Project Epsom gave all assessments to the same people, which is unusual.

Is 29 different assessments a lot to assess?
Yes. Project Epsom was a co-validation across assessments. It offers a more powerful comparison of different questionnaires. The assessments ranged from 10 minutes to one hour. The order of the assessments was varied and questionnaires went out over a period of over two months. This reduced fatigue from completing too many assessments in one sitting.

Were the participants paid?
Yes. Research of this scale and nature requires incentivising the participants. They had the option of donating their payment to charity, which many did.

Was a single-item measure of overall effectiveness used?
No. A three-item measure of overall effectiveness was used with good internal consistency, r=.67 (N=308).

Is the research in line with contemporary psychological theory and other research?
Yes. This research is consistent with worldwide meta-analytic research on how personality correlates with work performance. It builds on new models of work performance (Kurz & Bartram, 2002; MacIver et al., 2006) and the higher order factors of personality identified by Digman (1997) and Musek (2007). In Project Epsom, the personality scales which were expected to correlate with specific performance criteria were stipulated before analysis. This is essential in science and reduces capitalisation on chance. Multiple regression or canonical analysis, techniques commonly used in other studies, lack such clear rationale. They are more likely to capitalise on chance effects and so can produce unrealistic results.

Can one study of some 308 participants offer conclusive proof of improved validity?
No. In scientific research, little, if anything, can be conclusively "proven". However, we can research which test is likely to be more valid. We can then place some statistical confidence in this assertion. Project Epsom shows differences between some assessments which are statistically significant.

Are you presenting more information on this research?
New Technical Manuals and User Guides for Wave Styles and the Saville PQ are being released in 2009. These provide more detailed information than is appropriate for this paper. Correlations between the different assessments and work performance criteria are included.

Further research is due to be published in peer reviewed articles, and presented in papers and symposia in 2009. Papers have already been accepted at the following conferences:

- The British Psychological Society's Division of Occupational Psychology Conference, Blackpool, UK, January
- The Society for Industrial and Organizational Psychology, New Orleans, USA, April
- The European Congress of Work Psychology and Organizational Psychology, Santiago de Compostela, Spain, May

Are you doing further research?
Yes. Further analysis and research is ongoing with new samples and participants. For example, differences between specific job groups are being investigated.