An Overview of the
Development, Validity and Reliability of
Version 3.0 of the English Insights Discovery Evaluator

Dr. Stephen Benton
Director of The Business Psychology Centre (bpc),
The University of Westminster, 309 Regent Street, London, W1B 2UW, UK

Dr. Corine van Erkom Schurink
Director of The Analytical Research Bureau (Pty) Ltd,
16 Central Avenue, Pinelands 7405, South Africa

Stewart Desson
Head of Learning, Insights Learning and Development Ltd,
Jack Martin Way, Dundee, DD4 9FF, Scotland, UK

This executive summary sets out the evidence for the Insights Discovery model's psychometric measurement of the four colours being both valid and reliable. It draws upon an extensive research and development programme undertaken between Insights Learning and Development Ltd. and the University of Westminster, aimed at the development of a psychometrically robust evaluator. The full methodological and statistical account of this programme may be found in technical papers produced at the University of Westminster's Business Psychology Centre (bpc). Psychometric assessment is a science based on 'objectively' measuring characteristics of human behavior and specifically here, personality. In order to do so, psychological questionnaire measures must meet certain criteria to be considered objective measures. This paper sets out to explain these psychometric criteria in easily understood terms. It is the authors' intention to make the statistics and arguments presented understandable to two different professional groups that may have a need to work with the Insights Discovery Evaluator (IDE). Firstly, professional psychometricians *and* secondly the wider community of Human Resource professionals. Four categories of information are presented covering 'item analysis', 'norms data', 'reliability' and 'validity'. Key statistics have been computed for each of these four areas and they have been benchmarked against international standards. The paper concludes that the measurement of the four colours is both valid and reliable.

Introduction

In 1998 Andi Lothian collaborated with Jeff Davis at the University of Westminster to develop the first version of the Insights Discovery model. This work formed a core component of Jeff Davis's PhD entitled "Jung's Typology – The Development of a Psychometric Tool". Dr. Stephen Benton, one of the authors of this paper, supervised the PhD. Since then, over 10 postgraduate dissertations supervised through the Business Psychology Centre, at the University of Westminster, have further developed the model and its applications. Any evaluator claiming to be a psychometric must meet the international standards clearly defined by both the American Psychological Association (APA) and the British Psychological Society (BPS). This paper presents strong evidence in support of the Insights Discovery Evaluator's (IDE) claim to be a high quality psychometric, that meets these standards. To convey these standards in easily understood terms, the key statistics needed to establish the psychometric qualities have been presented in a four segment pyramid.

Peer Review Acknowledgements

We would like to thank the following people for their critiquing of this paper in its early drafts and their invaluable contribution to its development.

******PAPER IS WITH REVIEWERS NOW ... This section still to be written based on peer review group feedback ******

Objectives

- To explain how the Insights Discovery model has been developed
- To present the evidence for the Insights Discovery Evaluator's (IDE) psychometric measurement of the four colours being both valid and reliable
- To benchmark this evidence against other comparable personality based psychometrics
- To present a high level summary of the case for the evaluator meeting the psychometric standards set out by both the American Psychological Association and the British Psychological Society

Pyramid of Key Psychometric Statistics

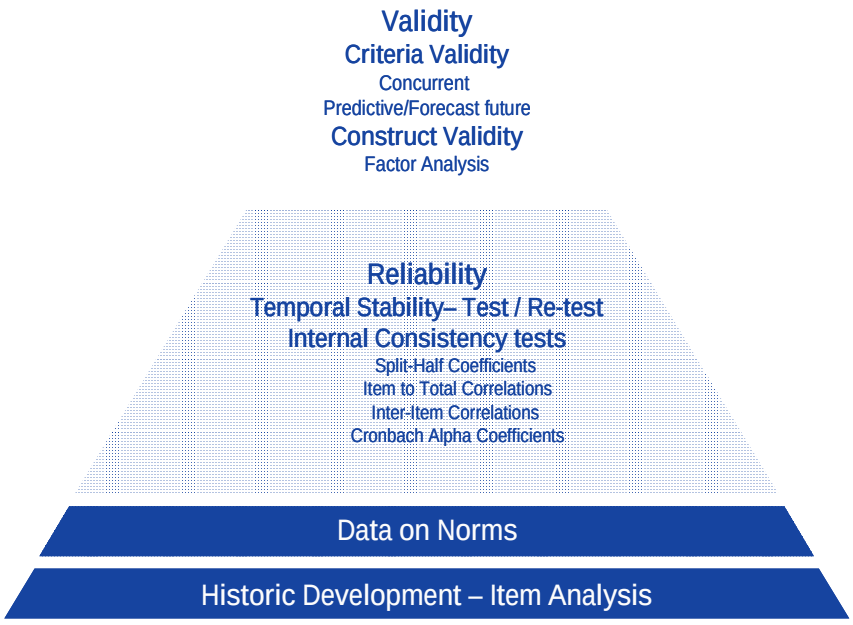


Figure One –Pyramid of Key Psychometric Statistics

Methodology

This paper focuses on the psychometric qualities of the Insights Discovery Evaluator (IDE) through analysis of the following samples:

- Sample sizes of between 350 and 2,000 have been used for three separate 'four colour item analysis' conducted between 1998 and 2005, resulting in the current English version 3.0 of the evaluator.
- The norms data presented here is a small subset of an analysis of 186,951 evaluators completed between 1/3/2000 and 31/7/2004. This includes evaluator data from the earlier versions 2.0 and 2.2 that have been improved through item analysis to become version 3.0. .
- The internal consistency reliability statistics are based on 24,224 version 3.0 evaluators completed between 31/11/2003 and 31/7/2004.
- The test/re-test reliability statistics are based on 1,435 evaluators completed between 1/5/2002 and 1/7/2004.
- The construct validity data is based on factor analysis using a sample of 7,159 version 3.0 evaluators completed between 1/2/2004 and 1/7/2004 and smaller sample sizes ranging between 1,259 and 3,425 based on Dutch, German and Canadian French and 'French French' Discovery translations of the evaluators.

This data has been collected from people completing evaluators in connection with them experiencing an Insights Discovery workshop or coaching session i.e. the context for doing the evaluator is developmental.

This paper draws on the APA's (American Psychology Association) book entitled 'Standards for Educational and Psychological Testing' (1999) as an authoritative source detailing the objective standards that all psychometrics must meet. In addition, two of Paul Kline's seminal texts entitled the 'Handbook of Psychological Testing' (2000) and the 'Psychometric Primer' (1997) have been used to define key psychometric concepts and as a source of key benchmark statistics for other comparable psychometrics.

The Insights Discovery Evaluator

In appendix A is version 3.0 of the English Insights Discovery evaluator. It is an ipsative (forced choice) and normative (a range) evaluator consisting of 25 frames in which the user chooses from a choice of four word pairs a 'most', a 'least' and then scores the remaining two options in between least and most on a scale of 1 to 5. Each of the 4 items in a frame measure preferences called 'Fiery Red', 'Sunshine Yellow', 'Earth Green' and 'Cool Blue'.

A completed evaluator will have 25 colour preference scores, each given a score between 6 (for most) and 0 (for least), for each of the four colours. A simple arithmetic mean across all 25 frames is calculated for each of the four colours. Figure two shows an example of the first 5 frames and an example of the colour chart produced from all 25 frames.

Input	**->**	**Output**
Example of 24 items in 6 frames	**->**	**Example of colour graph**

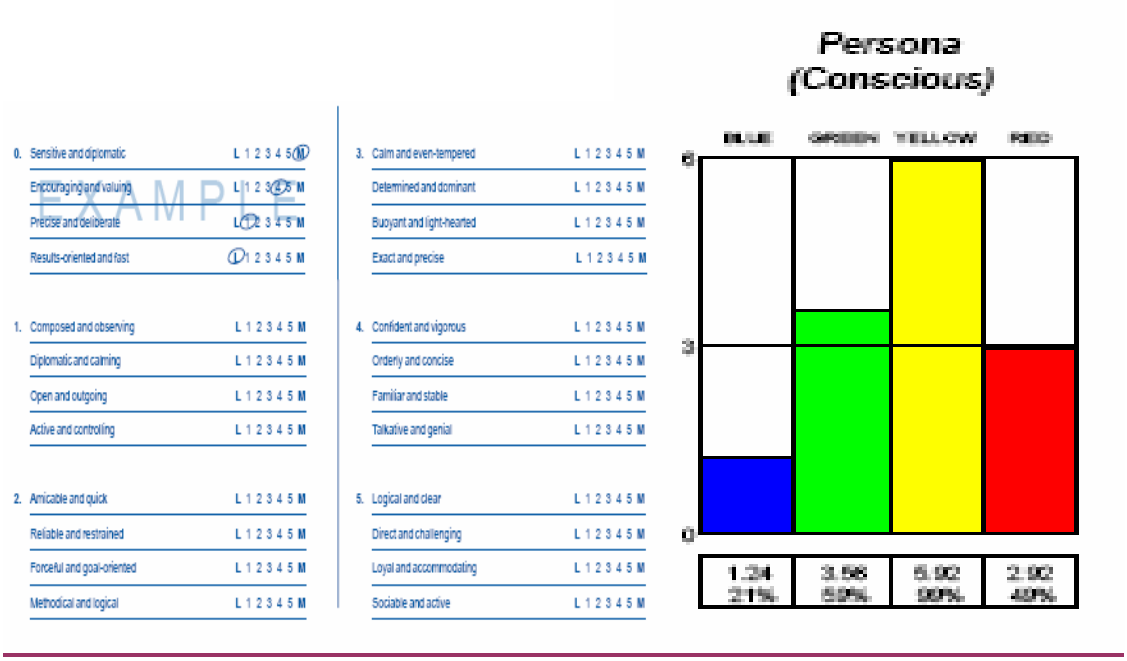


Figure Two – Sample of evaluator frames and example of profile output

Historic Development – Item Analysis

One of the great contributors to psychometrics, the late Professor Paul Kline (Psychometrics Primer, 1997), described this technique and supported the item analytic approach, saying:

> "Item Analysis is a simple and effective method of test construction and many well-known tests have been developed using this approach"

Page 39 of the APA's (American Psychology Association) 'Standards for Educational and Psychological Testing' explains "The test developer usually assembles an item pool that consists of a larger set of items that what is required by the test specifications. This allows for the test developer to select a set of items for the test that meet the test specifications. The quality of the items is usually ascertained through item review procedures and pilot testing". An example of the Insights Discovery Evaluator's (IDE) item analysis is provided in this paper and full documentation can be found in other technical papers produced at the University of Westminster's Business Psychology Centre (bpc).

One of the core methods underpinning the development of the IDE has been the test building technique of 'item analysis'. There are 100 colour 'items' (i.e. questions) spread over the 25 frames in the IDE. Item analysis involves empirically testing the quality of these 100 items and replacing weaker items with better ones. An example of a Fiery Red colour item from the evaluator is *'determined and resolute'*.

Item analysis has been used to produce from a wide pool of items, four colour based sub-sets of 25 items that are homogeneous, internally consistent and univariate within a colour i.e. each word pair statement measures just *one* colour.

One example of an item analysis on the 25 Sunshine Yellow items is show in the diagram below. On the horizontal axis are the 25 frames. On the vertical axis are the number of respondents that highlighted Sunshine Yellow, Fiery Red, Earth Green or Cool Blue as the 'most' in the evaluator. However, this sample of respondents are all people who have scored 5 or more (out of 6) for Sunshine Yellow. Consequently, we would expect that the Sunshine Yellow line should always be significantly above the other three coloured lines. It can be seen that the 6th frame of the questionnaire is weak. We would expect the people in this sample to consistently select the Sunshine Yellow item ahead of the other three colours. However, graph at the top of figure three shows that in the 6th frame they selected the Earth Green item (*accommodating and helping*) ahead of the Sunshine Yellow item (*upbeat and hopeful*).

The 6th frame was therefore subjected to systematic variation and re evaluation as new word pairs were empirically tested. The best results were found when the Earth Green item was changed from *'accommodating and helping'* to *'relating and amenable'*, combined with a change in the Sunshine Yellow item from *'upbeat and hopeful'* to *'expressive and hopeful'*.

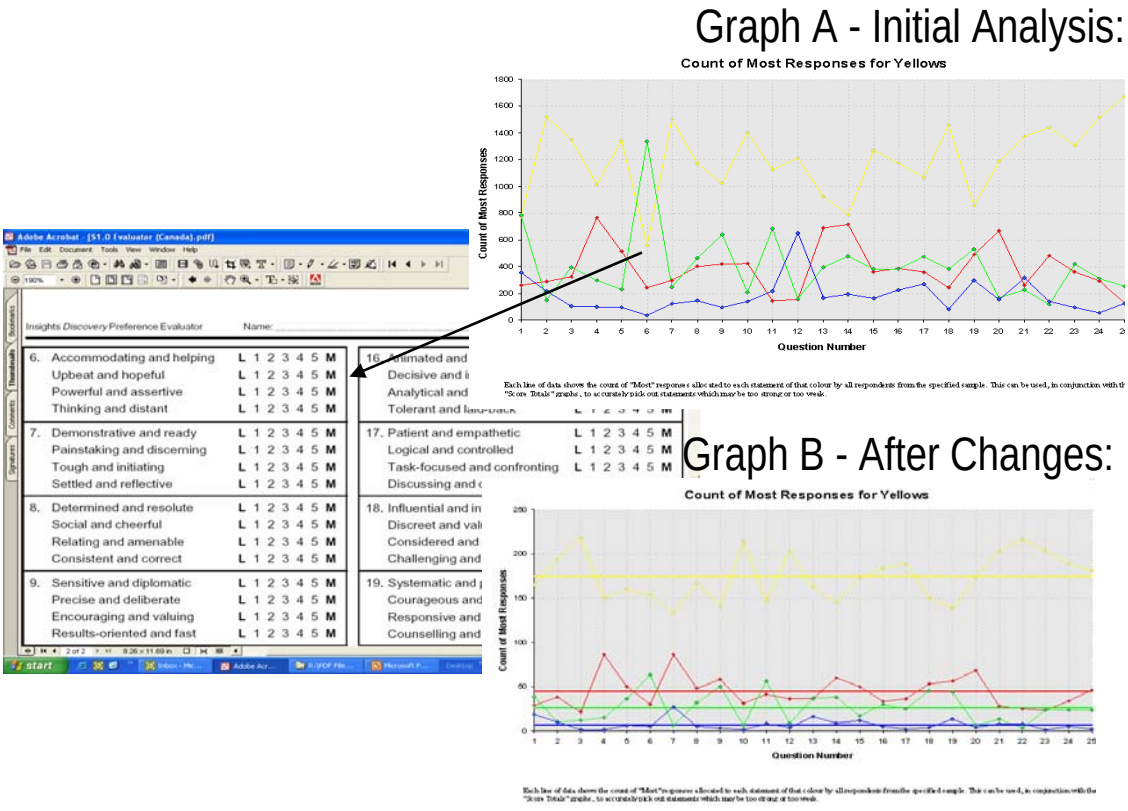


Figure Three – Item analysis Graphs before and after item changes

This procedure was repeated for all 25 frames and when the empirical results showed a better word pair, it replaced the weaker one. See the graph at the bottom of figure three for the same statistics produced after a series of word pair improvements across the evaluator.

Consequently the quality of the evaluator has been systematically improved over time through item analysis.

This data is based on 5,645 evaluators completed in July 2004.

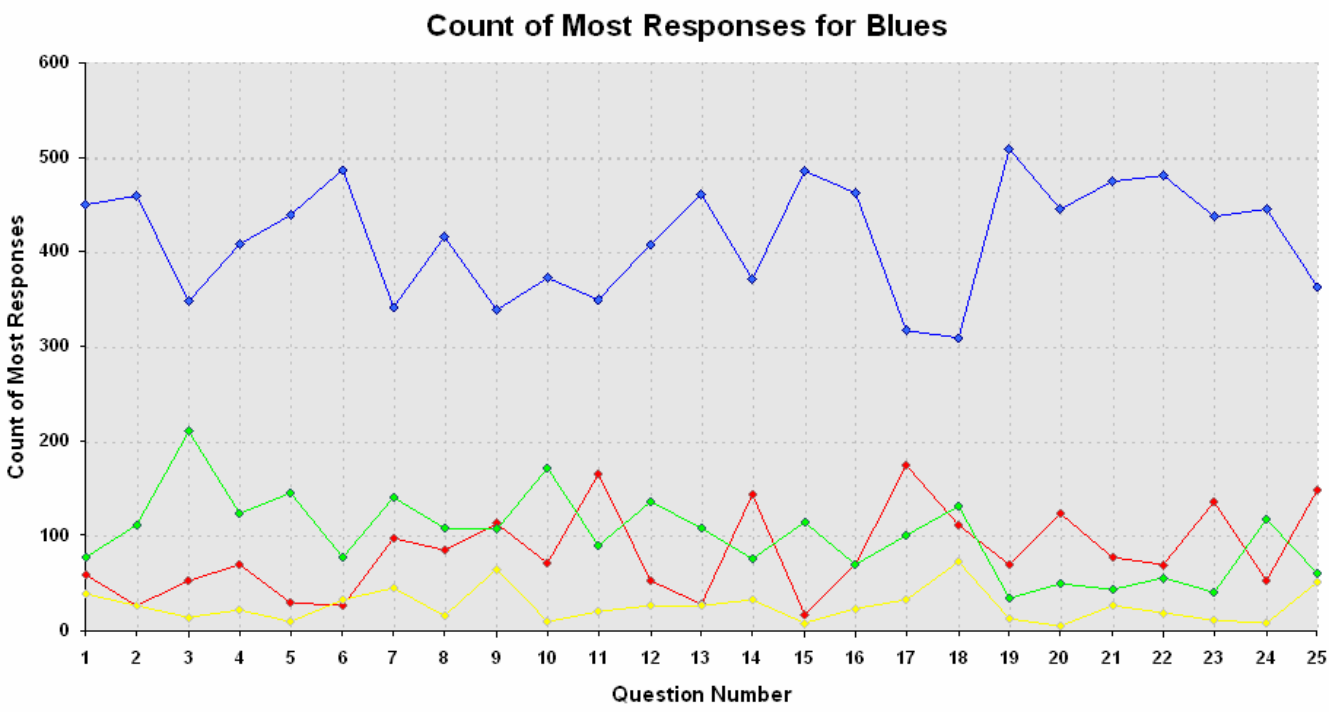


Figure Four – Cool Blue Item Analysis

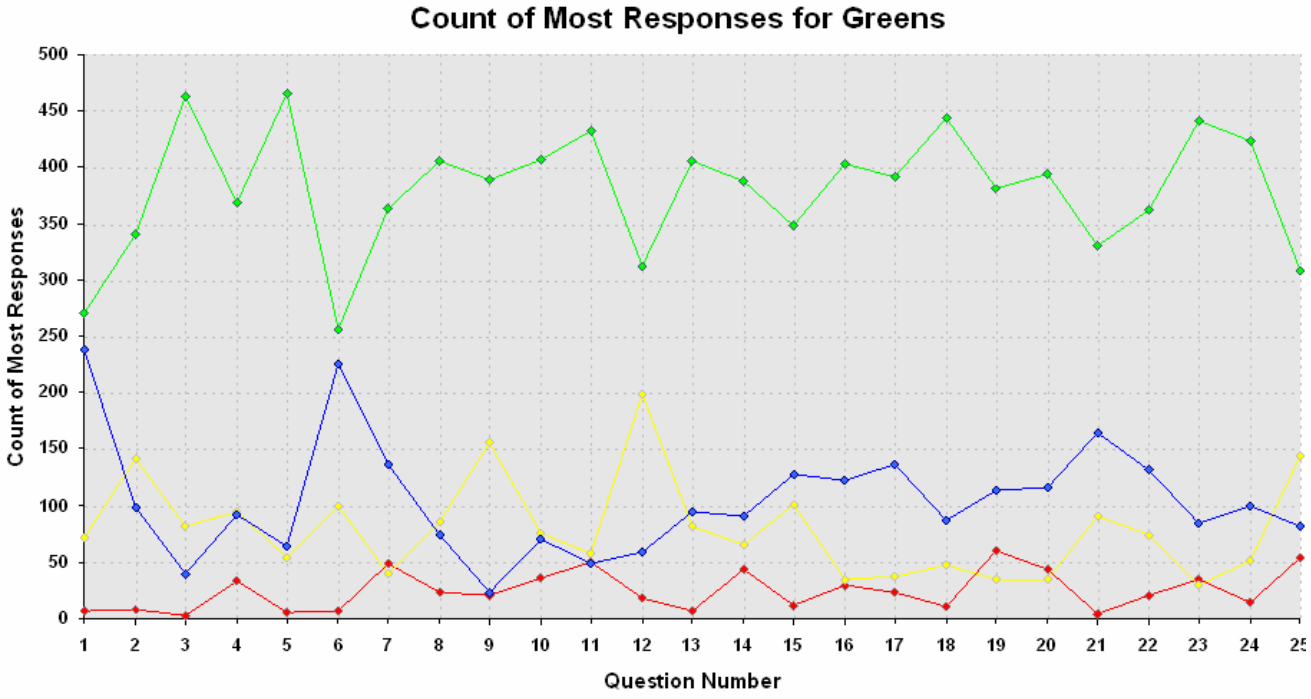


Figure Five – Earth Green Item Analysis

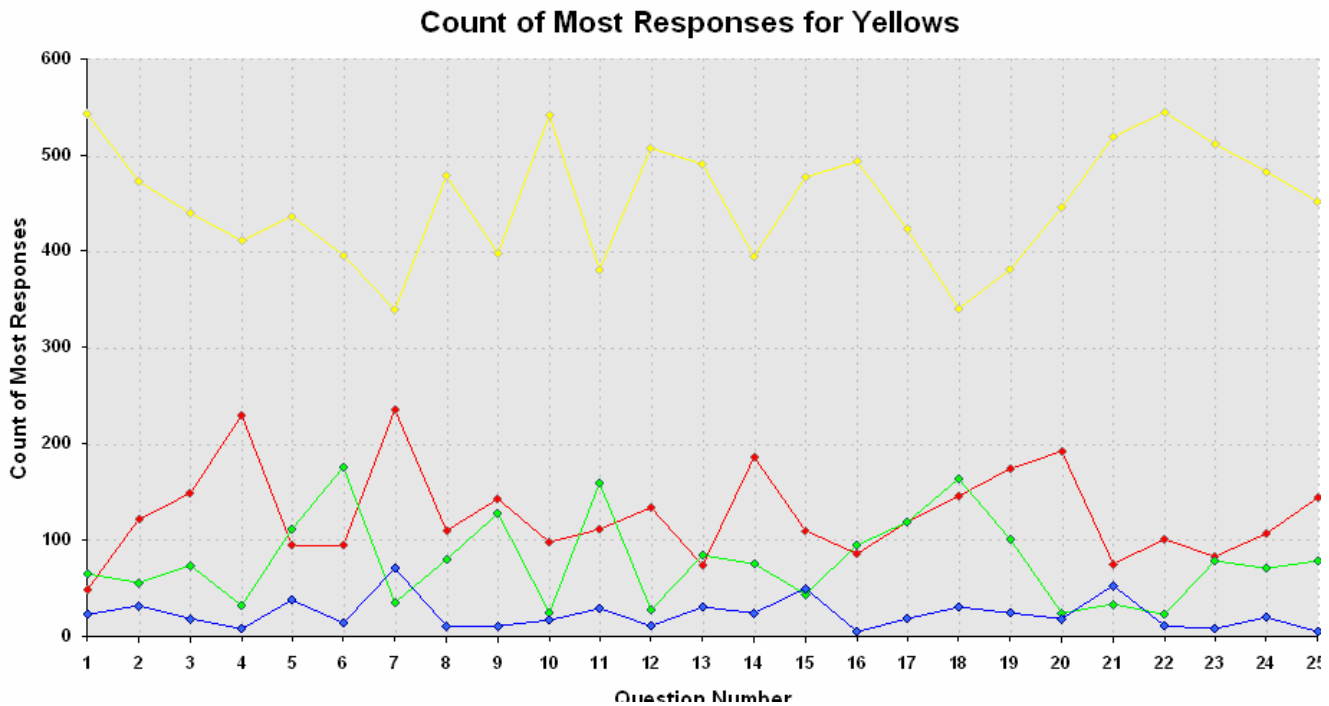


Figure Six – Sunshine Yellow Item Analysis

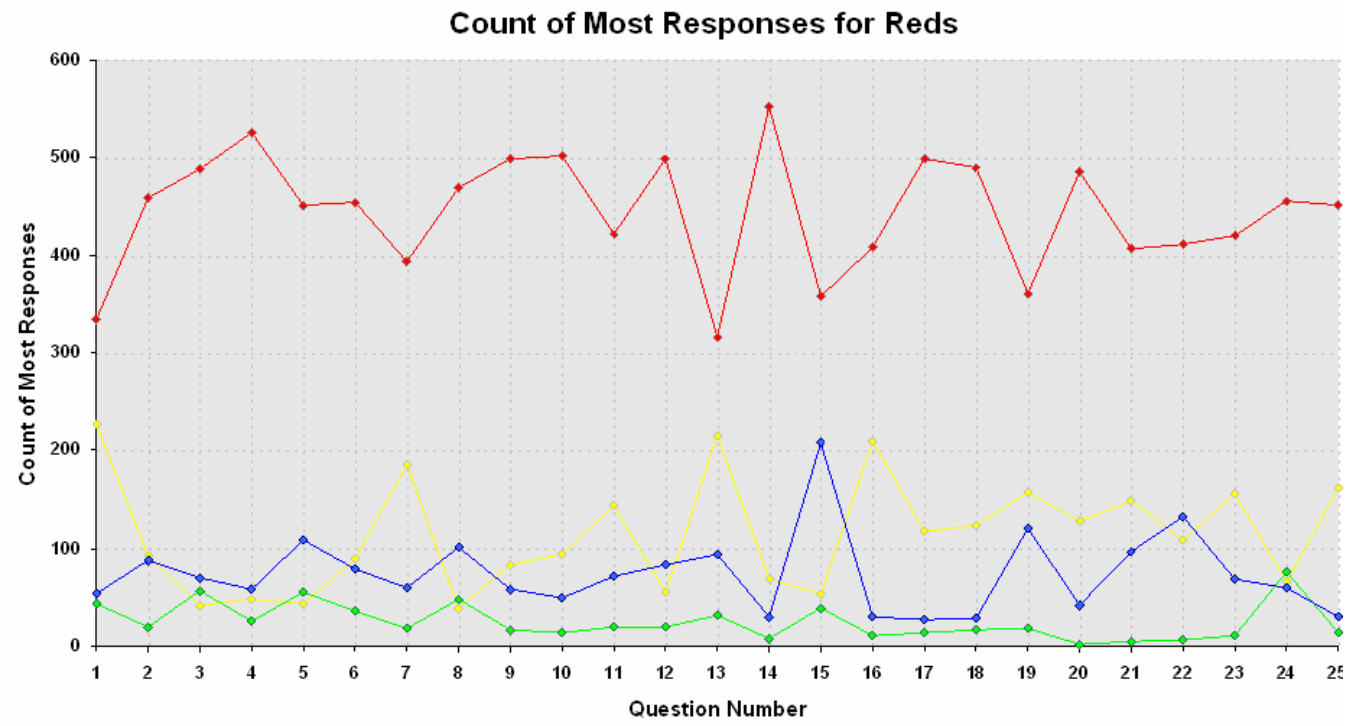


Figure Seven – Fiery Red Item Analysis

Developing the evaluator using item analysis has significantly improved the reliability of the evaluator. Figures four, five, six and seven graphically show the item analysis has evolved the evaluator to a very solid position.

Although the graphs presented here are both visually impactful and intuitively appealing, there is a need for a more statistical approach to quantity the quality of the item analysis. Consequently, t-tests have been conducted on the item analysis data to analyze statistically the graphical data. A t-test allows you to determine if the distance between the two colour scores within one frame, are statistically significant or not. Below is an example for the cool blue items from 4 samples comparing data across 4 continuously improving versions of the evaluators. This English Version 3.0 sample has been taken between 31/11/2003 and 31/7/2004.

		UK - English R25		UK - English S2		UK - English S2.2		UK - English S3.0	
Pair tested	Blue score>4 VS other colors	t value on paired t-test	Sig. of t (2-tailed) CI=95%	t value on paired t-test	Sig. of t (2-tailed) CI=95%	t value on paired t-test	Sig. of t (2-tailed) CI=95%	t value on paired t-test	Sig. of t (2-tailed) CI=95%
Pair 7	B3 & G3	3.11	0.002	2.73	0.006	11.85	0.000	9.53	0.000
Pair 19	B7 & G7	7.44	0.000	6.00	0.000	23.99	0.000	40.73	0.000
Pair 22	B8 & G8	14.40	0.000	10.53	0.000	25.80	0.000	26.51	0.000
Pair 25	B9 & G9	6.29	0.000	3.59	0.000	9.23	0.000	11.56	0.000
Pair 26	B9 & Y9	10.10	0.000	7.12	0.000	18.40	0.000	11.78	0.000
Pair 28	B10 & G10	11.85	0.000	9.33	0.000	32.73	0.000	36.82	0.000
Pair 37	B13 & G13	-8.71	0.000	22.98	0.000	26.45	0.000	27.99	0.000
Pair 49	B17 & G17	13.20	0.000	12.60	0.000	33.10	0.000	28.01	0.000
Pair 51	B17 & R17	10.27	0.000	15.67	0.000	38.39	0.000	39.75	0.000
Pair 52	B18 & G18	-1.09	0.276	-1.20	0.232	-5.89	0.000	10.47	0.000
Pair 70	B24 & G24	8.84	0.000	13.91	0.000	33.26	0.000	28.90	0.000

POTENTIAL PROBLEM: The two items (t-test) are very close (low positive t-value), but the distance between the two items is still significant

PROBLEM: The two items (t-test) are very close (low positive/negative t-value), and the distance between the two items is NOT significant

PROBLEM: The two items (t-test) are in inverted order (negative t-value), and the distance between the two items is/is not significant

Figure Eight - Table of t-tests on item analysis data

However, even a high quality item analysis does not necessarily ensure an evaluator is valid. Consequently, quoting further from Paul Kline in his Psychometrics Primer (1997), “After the items have been selected by item analysis and the results replicated with a new sample, it is necessary, as has been argued, to show the test is valid and reliable”.

Data on Norms

BPS (British Psychological Society) and APA (American Psychological Association) standards state that all psychometrics must supply norms for comparative purposes, the form of which varies according to the constructs and attributes being measured. Norms must be up to date and appropriate for the intended usage and population.

The norms data for the IDE is of good quality, being segmented by occupation (over 35 different occupations analysed), gender, age (in ten year bands) and the language of the evaluator (over 22 languages analysed).

However, care must be taken when making use of this data, so as not to make invalid statistical interpretations. For many psychometrics, norms are used as a reference against which an individual's psychological test results can be interpreted relative to a larger population. For example, a test of IQ might produce a score of 150. However, without a large and relevant sample population to provide a spread or distribution of scores against which to interpret the individual's score, the number 150 is without value. However, if we know a score of 150 places the individual in the top 5% of the distribution of scores for people of this age, from a certain socio-economic group, then the information becomes more useful.

Similarly, in clinical and related tests measuring a condition, norms provide a key reference for interpreting individual's scores and responses. Here it is important to establish if an individual's score, relative to the population, is indicative of 'more' or 'less' of the construct in question. For example; more depressed, higher IQ, slower reactions time or faster and/or more accurate short-term memory recall. Norms, for these forms of tests, provide a means of assessing a person's relative standing in comparison to others.

However, being able to measure whether or not someone has more or less personality isn't the aim of the IDE. The norms data presented in this paper should *not* be used to make comparisons between an individual's colour scores and the continuous population distribution. One reason this comparison should not be made is that the evaluator is 'ipsative' and involves a 'forced choice' that results in the four colour energies being ranked 25 times. It is not statistically valid to treat ranked data as normative and relate it to a continuous distribution from a wider population i.e. to produce statistics on where (say) an individual's aggregate 'Fiery Red' score sits within a population's continuous distribution curve would be in statistically erroneous. For a fuller explanation of this argument please see page 49 in Professor Paul Kline's 'Psychometric Primer' (1997).

Instead, for the purposes of personality profiling derived from an 'ipsative' personality preference evaluator, norms data is typically used to compare an individual's dominant preference with the percentage of the population that have the same dominant preference. This is typically examined across ages, professions and cultures.

A sample of the data is given below for people speaking French in both Canada (on the left) and France (on the right). This data is not a *random* sample of the population using these evaluators in different languages, but a *convenient* sample of all those participants that have (for whatever reason), experienced an Insights Discovery workshop or coaching session.

A French Canadian whose individual results show a dominant preference for Fiery Red, may be interested to know that just over 22% of the population also has the same dominant preference, while 78% have a preference for a different colour (see the bottom left hand graphs below).

This example is just a small sample of a very large set of norms data available in the more detailed papers. The norms data provides good evidence of the 'predictive validity' of the model (see the later section in this paper on criteria validity for a fuller exploration of this). Predictive validity indicates a model can 'predict' something. For the Insights Discovery model, we can predict that certain professions are likely to have a higher percentage of a dominant colour energy. Accountants, for example, are more likely to have a preference for blue energy. This does *not* mean that to be a successful accountant, you must have a dominant blue energy; we do not measure *capability* in the Discovery model, we have only measured *preference.*

Although this norms data provides good evidence of the 'predictive validity' of the model, it does ***not*** imply the evaluator is valid for use in recruitment. The data below only presents aggregate data on *preferences* and says nothing about the sample's *capabilities*. If a practitioner of the Discovery system were to use this norms data as the basis for recruitment to predict who the capable candidates may be, this would be an unethical and discriminatory practice. However, in the authors' experience this is a common error made by inept practitioners working with other psychometric tools on the market. This error should not be made with the IDE.

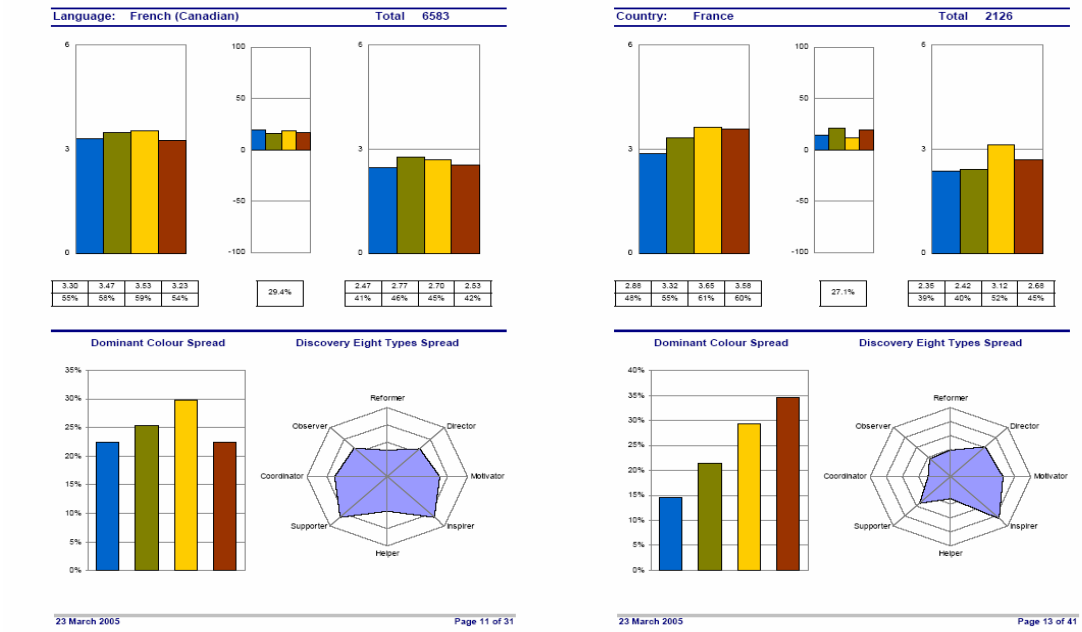


Figure Nine – Example of Norms data for French speakers in Canada and France

Below is a tabular display for the norms data for all the languages analysed.

The table shows one row for each population segment.

Population Segment	Sample Size	Average Colour Scores				Percentage of population with dominant colour energy			
		Blue	Green	Yellow	Red	Blue	Green	Yellow	Red
Australia	1175	3.45	3.50	3.47	3.13	27.40%	26.13%	25.19%	21.28%
Austria	368	3.18	3.35	3.60	3.31	24.18%	26.09%	26.36%	23.37%
Belgium	357	3.24	3.36	3.40	3.56	21.29%	23.53%	24.65%	30.53%
Brazil	221	3.01	3.33	3.41	3.97	14.03%	12.67%	27.60%	45.70%
Canada	30171	3.36	3.64	3.46	3.10	23.87%	29.60%	25.90%	20.63%
Denmark	230	2.74	3.43	3.60	3.78	5.65%	28.26%	24.78%	41.30%
Finland	583	3.42	3.12	3.14	3.76	21.10%	20.93%	22.47%	35.51%
France	2126	2.88	3.32	3.65	3.58	14.63%	21.40%	29.40%	34.57%
Germany	2601	3.22	3.32	3.50	3.34	23.03%	23.30%	27.22%	26.45%
Hong Kong	128	3.81	3.82	3.20	2.90	31.25%	34.38%	17.97%	16.41%
Hungary	178	3.63	3.59	3.48	3.03	29.21%	25.84%	23.60%	21.35%
Ireland	760	3.33	3.52	3.55	3.13	24.47%	27.11%	26.05%	22.37%
Italy	411	3.00	3.11	3.67	3.70	17.03%	15.33%	30.66%	36.98%
Mexico	2020	3.39	3.53	3.16	3.65	23.02%	27.57%	17.87%	31.53%
N Ireland	100	3.09	3.39	3.78	3.32	15.00%	24.00%	30.00%	31.00%
Netherlands	1692	2.75	3.36	3.51	3.94	12.65%	21.16%	23.82%	42.38%
Poland	427	3.38	3.51	3.04	3.43	24.36%	29.04%	16.63%	29.98%
Singapore	221	3.70	3.37	3.01	3.42	36.20%	22.17%	13.57%	28.05%
South Africa	2691	3.43	3.41	3.30	3.54	22.78%	27.02%	19.06%	31.14%
Spain	681	3.14	3.69	3.52	3.27	17.62%	28.49%	25.70%	28.19%
Sweden	154	3.10	3.67	3.43	3.36	15.58%	38.96%	17.53%	27.92%
Switzerland	2865	3.07	3.22	3.65	3.58	20.39%	19.68%	29.17%	30.76%
Turkey	377	3.83	3.20	3.26	3.67	34.75%	16.71%	23.08%	25.46%
UAE	141	3.23	3.34	3.43	3.76	23.40%	18.44%	21.99%	36.17%
UK	33678	3.22	3.40	3.65	3.18	22.98%	24.20%	29.42%	23.40%
USA	40570	3.40	3.61	3.41	3.21	23.77%	29.14%	24.47%	22.62%

Figure Ten - A tabular summary of the language norms data

****** Lindsay to update figure ten with LANGAUGE DATA (THIS IS CURRENTLY COUNTRY DATA. WE DO NOT WANT COUNTRY DATA IN THE PAPER) ******

****** Lindsay to update figures eleven and twelve with data with actual information ******

Population Segment	Sample Size	Average Colour Scores				Percentage of population with dominant colour energy			
		Blue	Green	Yellow	Red	Blue	Green	Yellow	Red
Males	6,583	55%	58%	59%	54%	22%	25%	30%	22%
Females	2,126	48%	55%	61%	60%	15%	21%	29%	35%

Figure Eleven - A tabular summary of the gender norms data

Population Segment	Sample Size	Average Colour Scores				Percentage of population with dominant colour energy			
		Blue	Green	Yellow	Red	Blue	Green	Yellow	Red
Age 10-19	6,583	55%	58%	59%	54%	22%	25%	30%	22%
Age 20-29	2,126	48%	55%	61%	60%	15%	21%	29%	35%
Age 30-31									
Etc.									

Figure Twelve - A tabular summary of the age band norms data

Overview of Reliability

Reliability has two meanings[1]:

"Does each item in the evaluator perform consistently?"
(called *Internal Consistency*)
"Do we have consistent results over a period of time?"
(called *Temporal Stability*)

A highly reliable evaluator will produce consistent colour scores that are repeatable over time. The ideal is to have high internal consistency and high temporal stability. However, all measurement procedures have the potential for error that reduces the reliability. The aim is to identify where the error is coming from (e.g. an unclear question) and to minimise it. Any observed score is made up of what statisticians call the *'true score'* plus the measurement of unwanted and/or unknown factors i.e. *'measurement error'*. In estimating the reliability of the Insights Discovery evaluator, we need to determine how much of the variability in the colour scores is due to *measurement error* and how much is due to real variability in the *true scores*. Measurement errors are essentially random: a person's colour scores might not reflect the true score because of a whole raft of reasons such as; they were sick, hung over, anxious, bored or trying to give answers they think would suit the expectations of others.

Reliability: Internal Consistency

Internal Consistency applies to the consistency of the scores amongst the 25 colour items i.e. it deals with measures of homogeneity within the colour items. The rationale for internal consistency is that the individual 25 colour items should all be measuring the same construct and thus be highly inter-correlated (Churchill,1979; Nunnally, 1979). Four types of internal consistency have been examined; inter-item reliability; item to total reliability; cronbach alpha reliability and split–half reliability.

'Inter-Item' and 'Item to Total' Reliability

'Inter-item' and 'item to total' correlations have been calculated using the Pearson Product-Moment Correlation. This involved creating four colour based '25 by 25' matrices showing the correlation between the 25 colour items. In addition, we have computed 'item-to-total' correlations by correlating the individual colour item score to the sum of all 25 scores for the same colour. An example of these correlations for the Cool Blue items on the evaluator is shown on the next page, with a summary table of statistics on the page after. In 1991, Robinson et al concluded that the mean 'inter-item' correlation should equal or exceed 0.30 for this to be good evidence of reliability. For 'item to total' correlations Robinson et al (1991) concluded that the correlation should equal or exceed 0.50 for this to be good evidence of reliability. The analysis of the 24,224 evaluators shows that, for each of the four colours in the evaluator, the average 'inter-item' correlation is significantly above 0.3 and the 'item to total' correlation is significantly above 0.5, providing strong evidence of the case for reliability.

[1] Reliability meanings based on pages 26 to 33 in 'A Psychometrics Primer' by Paul Kline (1997)

University of Westminster

bpc

Figure Thirteen Correlation Coefficients – Insights Discovery evaluator version S 3.0 (UK) N = 24,224

Cool Blue color preference																											
Frame & statement code	**Item-Total Correlation Coefficients**	Squared Item-Item Correlation Coefficients	1_1	2_4	3_4	4_2	5_1	6_4	7_2	8_4	9_2	10_4	11_2	12_1	13_3	14_2	15_1	16_3	17_2	18_3	19_1	20_3	21_1	22_3	23_4	24_3	25_1
1_1	0.45	0.29	1.00	0.23	0.19	0.26	0.27	0.43	0.24	0.22	0.17	0.22	0.22	0.22	0.26	0.28	0.20	0.36	0.37	0.29	0.26	0.30	0.24	0.23	0.30	0.36	0.31
2_4	0.56	0.37	0.23	1.00	0.39	0.36	0.46	0.30	0.18	0.32	0.32	0.38	0.37	0.35	0.42	0.35	0.40	0.36	0.25	0.26	0.39	0.36	0.27	0.34	0.36	0.35	0.30
3_4	0.59	0.45	0.19	0.39	1.00	0.47	0.32	0.26	0.12	0.43	0.44	0.56	0.40	0.38	0.34	0.30	0.35	0.34	0.21	0.20	0.40	0.32	0.33	0.37	0.49	0.36	0.40
4_2	0.60	0.43	0.26	0.36	0.47	1.00	0.36	0.31	0.14	0.41	0.36	0.49	0.33	0.33	0.31	0.31	0.38	0.35	0.29	0.22	0.48	0.35	0.46	0.38	0.47	0.39	0.39
5_1	0.54	0.36	0.27	0.46	0.32	0.36	1.00	0.31	0.15	0.35	0.26	0.31	0.39	0.31	0.39	0.37	0.37	0.33	0.27	0.26	0.40	0.34	0.29	0.33	0.33	0.36	0.28
6_4	0.53	0.35	0.43	0.30	0.26	0.31	0.31	1.00	0.24	0.30	0.25	0.26	0.27	0.29	0.35	0.33	0.26	0.40	0.40	0.31	0.30	0.37	0.26	0.31	0.33	0.38	0.36
7_2	0.29	0.16	0.24	0.18	0.12	0.14	0.15	0.24	1.00	0.11	0.10	0.14	0.13	0.21	0.26	0.18	0.13	0.27	0.14	0.21	0.13	0.23	0.13	0.12	0.15	0.32	0.19
8_4	0.56	0.35	0.22	0.32	0.43	0.41	0.35	0.30	0.11	1.00	0.38	0.41	0.39	0.31	0.30	0.34	0.32	0.31	0.27	0.23	0.41	0.34	0.30	0.38	0.42	0.35	0.37
9_2	0.50	0.31	0.17	0.32	0.44	0.36	0.26	0.25	0.10	0.38	1.00	0.40	0.37	0.33	0.28	0.31	0.31	0.27	0.24	0.20	0.35	0.29	0.24	0.32	0.38	0.28	0.33
10_4	0.61	0.49	0.22	0.38	0.56	0.49	0.31	0.26	0.14	0.41	0.40	1.00	0.38	0.41	0.34	0.30	0.36	0.36	0.22	0.21	0.43	0.33	0.38	0.35	0.58	0.41	0.42
11_2	0.54	0.33	0.22	0.37	0.40	0.33	0.39	0.27	0.13	0.39	0.37	0.38	1.00	0.33	0.34	0.35	0.33	0.33	0.26	0.23	0.35	0.32	0.27	0.33	0.38	0.34	0.34
12_1	0.56	0.38	0.22	0.35	0.38	0.33	0.31	0.29	0.21	0.31	0.33	0.41	0.33	1.00	0.50	0.31	0.38	0.42	0.22	0.25	0.35	0.37	0.25	0.33	0.39	0.37	0.38
13_3	0.58	0.43	0.26	0.42	0.34	0.31	0.39	0.35	0.26	0.30	0.28	0.34	0.34	0.50	1.00	0.35	0.38	0.49	0.24	0.31	0.37	0.41	0.24	0.36	0.36	0.40	0.33
14_2	0.55	0.32	0.28	0.35	0.30	0.31	0.37	0.33	0.18	0.34	0.31	0.30	0.35	0.31	0.35	1.00	0.32	0.36	0.33	0.33	0.38	0.39	0.26	0.36	0.36	0.35	0.31
15_1	0.55	0.35	0.20	0.40	0.35	0.38	0.37	0.26	0.13	0.32	0.31	0.36	0.33	0.38	0.38	0.32	1.00	0.36	0.22	0.23	0.47	0.36	0.32	0.37	0.37	0.36	0.30
16_3	0.62	0.43	0.36	0.36	0.34	0.35	0.33	0.40	0.27	0.31	0.27	0.36	0.33	0.42	0.49	0.36	0.36	1.00	0.35	0.35	0.37	0.43	0.32	0.38	0.42	0.44	0.41
17_2	0.47	0.30	0.37	0.25	0.21	0.29	0.27	0.40	0.14	0.27	0.24	0.22	0.26	0.22	0.24	0.33	0.22	0.35	1.00	0.31	0.32	0.35	0.26	0.31	0.32	0.31	0.33
18_3	0.43	0.23	0.29	0.26	0.20	0.22	0.26	0.31	0.21	0.23	0.20	0.21	0.23	0.25	0.31	0.33	0.23	0.35	0.31	1.00	0.25	0.31	0.19	0.26	0.28	0.31	0.25
19_1	0.64	0.47	0.26	0.39	0.40	0.48	0.40	0.30	0.13	0.41	0.35	0.43	0.35	0.35	0.37	0.38	0.47	0.37	0.32	0.25	1.00	0.44	0.48	0.45	0.48	0.42	0.39
20_3	0.60	0.39	0.30	0.36	0.32	0.35	0.34	0.37	0.23	0.34	0.29	0.33	0.32	0.37	0.41	0.39	0.36	0.43	0.35	0.31	0.44	1.00	0.31	0.44	0.40	0.42	0.38
21_1	0.51	0.34	0.24	0.27	0.33	0.46	0.29	0.26	0.13	0.30	0.24	0.38	0.27	0.25	0.24	0.26	0.32	0.32	0.26	0.19	0.48	0.31	1.00	0.29	0.42	0.37	0.33
22_3	0.57	0.37	0.23	0.34	0.37	0.38	0.33	0.31	0.12	0.38	0.32	0.35	0.33	0.33	0.36	0.36	0.37	0.38	0.31	0.26	0.45	0.44	0.29	1.00	0.43	0.37	0.39
23_4	0.66	0.49	0.30	0.36	0.49	0.47	0.33	0.33	0.15	0.42	0.38	0.58	0.38	0.39	0.36	0.36	0.37	0.42	0.32	0.28	0.48	0.40	0.42	0.43	1.00	0.45	0.45
24_3	0.62	0.41	0.36	0.35	0.36	0.39	0.36	0.38	0.32	0.35	0.28	0.41	0.34	0.37	0.40	0.35	0.36	0.44	0.31	0.31	0.42	0.42	0.37	0.37	0.45	1.00	0.40
25_1	0.59	0.37	0.31	0.30	0.40	0.39	0.28	0.36	0.19	0.37	0.33	0.42	0.34	0.38	0.33	0.31	0.30	0.41	0.33	0.25	0.39	0.38	0.33	0.39	0.45	0.40	1.00

The covariance matrix is calculated and used in the analysis.

Correlation of items with themselves (perfect correlation)

Weak items ('item to total' correlation coefficient ≤0.50 as well as 'Item to Item ≤0.30)

Acceptable coefficients for 'item to total' correlation (≥0.50)

Acceptable coefficients for 'item to item' correlation (≥0.30)

For example, the Cool Blue item *'methodical and logical'* (frame 2, question 4) has a correlation coefficient of 0.36 with the item *'orderly and concise'* (frame 4, question 2). As this is above 0.3, it is considered a good result.

Below is a summary of the 'inter-item' correlations for this Cool Blue data and the other three colours. The top row of statistics in the table below show the average 'inter-item' correlations are significantly above 0.3. In addition, a high percentage of the colour items (between 18 and 22 out of 25) are statistically 'strong'.

N = 24,224 Inter-Item Correlations	**Colour preference**			
	Cool Blue	**Earth Green**	**Sunshine Yellow**	**Fiery Red**
Mean	**0.33**	**0.32**	**0.31**	**0.35**
Minimum	0.10	0.12	0.06	0.13
Maximum	0.58	0.60	0.60	0.56
Range	0.48	0.47	0.55	0.43
Maximum/Minimum	5.68	4.84	10.88	4.42
Variance	0.01	0.01	0.01	0.01
N of items in the scale	25	25	25	25
N of weak items N of strong items	3 22	3 22	7 18	3 22

Figure Fourteen – Inter-Item Correlations

Cronbach Alpha Reliability

In addition to the 'inter-item' and 'item to total' correlations, another important measure of reliability is the Cronbach Alpha coefficient. The coefficient measures the error variance on the average inter-item correlation. When the error variance is low, which is desirable, the alpha coefficient approaches 1.0. A value of 0.70 is the commonly accepted inferior limit (see DeVellis, 1991; Robinson & Shaver, 1973; Robinson & al, 1991; Swailes & McIntyre-Bhatty, 2002).

Analysing the same 24,224 completed evaluators shows the four colours have very high Cronbach Alpha coefficients, providing further evidence of excellent reliability.

N = 24,224	**Colour preference**			
	Cool Blue	**Earth Green**	**Sunshine Yellow**	**Fiery Red**
Cronbach Alpha Coefficients	**0.924**	**0.921**	**0.932**	**0.917**

Figure Fifteen – Cronbach Alpha Coefficients

Split Half Reliability

The final measure of Internal Consistency that supports the case for reliability is the 'split-half' measure. In split-half reliability we randomly divide all items that are thought to measure the same construct into two sets e.g. we create two sets of red items. We test the evaluator on a sample of people and compute the total score for each randomly divided half. The split-half reliability estimate is the correlation between these two total scores.

The split-half measures for the Insight evaluator were achieved by splitting the 25 frames into two groups of 12 and 13. The colour results are computed for each of the two groups and then correlated. High correlation suggest high reliability i.e. the higher the association (correlation coefficient) between the two data sub-sets, the higher the internal consistency of the scale. The results of a 'split-half' analysis also show high coefficients for the IDE:

- Cronbach Alpha Coefficients above 0.8 for each half
- Pearson Correlation Coefficients above 0.7 i.e. the 2 halves correlate highly

Reliability: Temporal Stability – Test / Re-test

Temporal Stability or Test-Re-test Reliability is determined through the administration of the same evaluator across time and it helps us gauge how robust the items are. If the results are statistically sound, then practitioners may have confidence in both the durability of results and their applicability across situations. This type of reliability is particularly useful for measures of stable personality traits, but not for measures of aptitude, where practice effects can significantly influence scores on future administrations.

There are 2 key reasons why an individual's re-test scores may differ from their original test. Firstly, there may be variability in their responses due to *measurement error* and therefore demonstrating a lack of reliability in the instrument. Secondly, they may have experienced personal change in this period and now genuinely have altered their colour scores. The Insights Discovery research team are continuously working to eliminate the first possible reason (instrument error). The research team is also working to understand and quantify the second reason in the belief that human beings are dynamic/ evolving. This approach acknowledges the possibility that people may shift their preferences numerous times over the course of a lifetime.

A *convenient sample* of 1,435 people who needed to complete the evaluator twice, had their original and re-tested colour scores assessed through a Pearson Correlation analysis. Reliability is expressed as correlation coefficients, ranging from 1 to 0. Temporal stability tests are generally expected to yield reliability coefficients ranging between 0.70 and 0.90. As a matter of comparison, studies published on other Jungian based instruments are reporting Correlation coefficients ranging from 0.69 to 0.83 (Carlson, 1985; Carlyn, 1977).

	RETEST Cool Blue Score	RETEST Earth Green Score	RETEST Sunshine Yellow Score	RETEST Fiery Red Score
TEST Cool Blue Score	**0.85**	0.12	-0.72	-0.30
TEST Earth Green Score	0.14	**0.81**	-0.17	-0.66
TEST Sunshine Yellow Score	-0.74	-0.16	**0.86**	0.15
TEST Fiery Red Score	-0.29	-0.65	0.13	**0.82**

Figure Sixteen – Test re-test correlation

The results of the Test/Retest analysis performed on the four colour scores show a very high reliability, translating into coefficients ranging from:

- 0.81 to 0.86 for the Pearsons Correlation Coefficients and
- 0.89 to 0.92 for the Cronbach Alpha Reliability Coefficients

As a matter of comparison, other studies published on the report the following range of coefficients:

- 0.69 to 0.83 - Correlation coefficients (Carlson, 1985; Carlyn, 1977 – on the MBTI instrument)
- 0.52 to 0.96 - Reliability Coefficients (Harrell & Lombardo, 1984 – on the 16PF)
- 0.76 to 0.84 - Reliability Coefficients (Capraro & Capraro, 2000 – on the MBTI).

The degree to which a test is reliable defines the accuracy with which it elicits and assesses someone's responses. However, just because a test is capable of delivering high reliability scores does not mean that it is valid i.e. a test, which is reliable, does not necessarily measure what it is supposed to. Just because a test is reliable and is consistent over time, does not mean that it is valid. An effective psychometric tool requires a combination of evidence supporting both its reliability and its validity.

Validity

Validity addresses *what* the evaluator actually measures and how well it measures it. For the IDE, validity is concerned with what can be interpreted from the colour scores. Psychometric measurements are always validated in regards to a particular use i.e. one cannot say that the evaluator has 'high' or 'low' validity per se. However, evidence can be gathered for interpretation of the colour scores being 'valid' in a particular way in which they are applied and used i.e. the context is always very important to any validity evidence presented.

Despite this, it is common for some well established tests to be erroneously referred to as having 'high validity' based on an unspoken assumption about how the test is used. All hidden assumptions must be made visible for a claim to validity to be authenticated.

In summary, validity means:

"Do we measure what we say we measure?"

Although this question may sound banal, providing psychometrically sound answers that meet international standards involves a substantial amount of work. The American Psychological Association 'Standards' publication says *'a sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses'.*

Different types of validity discussed in this section include:

- Face Validity - do the items (inputs) and/or the measures (outputs) from a test appear plausible to the user?
- Content Validity - can an expert objective source validate the quality of the items?
- Criteria Validity - predictive. This refers to the degree to which a test can predict a person's behaviours or performance on future, specified activities.
- Criteria Validity - concurrent. Here the validity of any test is best determined by comparing it to another test or some observable fact i.e. criteria validity is always based on external relationships.
- Construct Validity - the degree to which the test measures the underlying theoretical construct.

Face Validity

There are two different applications of the term face validity. The first concerns the degree to which the items in a test *appear* to measure what the test claims to measure. The second concerns the extent to which the users of the test *believe* the outputs from a test are accurate, as defined by how the outputs match their *self perception*.

Although it is usually considered desirable for a test's items to appear valid, this may not always be the case. For example, on measures geared toward the assessment of malingering and deception, low face validity may aid in more effective detection.

However, for personality tests such as the IDE, having the items appear valid is desirable in that it helps ensure users are willing to fill in the evaluator.

Without reasonable face validity, the users' confidence in a personality test may be undermined if they do not think the items are plausible when they are completing it or if they disagree strongly with the outputs.

However, although user confidence is important, it is not presentable evidence in making a psychometric claim to validity. One reason for this is the so called "Barnum Effect". In 1949 psychologist Bertram Forer gave a personality test to a group having told them the results were individually personalised. Unbeknown to the group, they actually were all given the same description based on an astrology book. The group then scored the accuracy of the results and the average score was 4.2 on a scale of 0 to 5, with over 40% claiming scoring it 5/5.

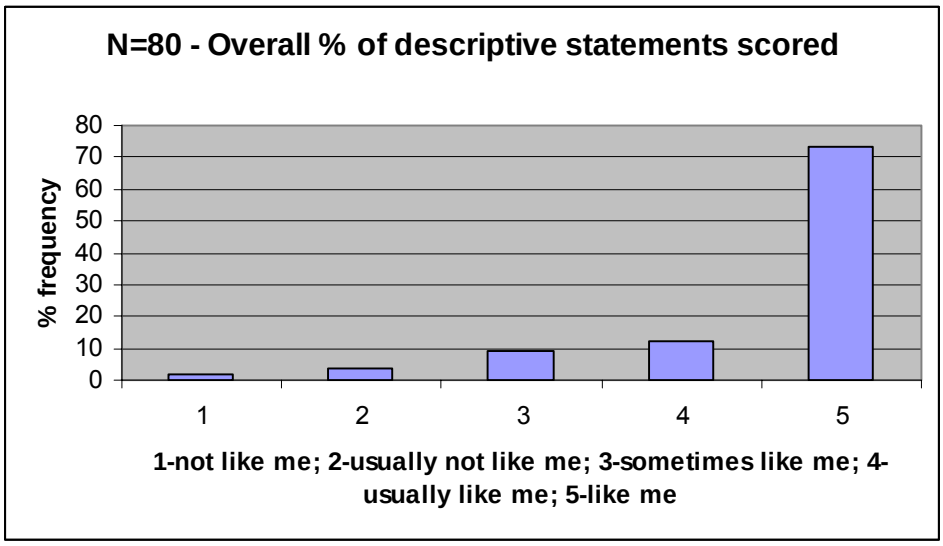
Despite face validity being of little relevance, some other instruments still offer face validity as the main evidence of their validity. This is not the IDE's main evidence. However, although having high face validity does not demonstrate overall validity, a lack of face validity would be a serious obstacle to practically using the evaluator. If user confidence is impacted by low face validity, people may choose not to use it or not to believe its output.

To provide you with confidence in the practical use of the IDE, here are some face validity results (with the caution that these statistics are not enough to establish overall validity).

In a University of Westminster survey (Remarczyk, 2005), a group of 80 people completed the IDE and were presented with their four colour scores accompanied by 50 sentences selected to describe the intensity of their personal four colour scores.

They were asked to mark out of 5 the overall accuracy of the information contained in the colour scores and the statements. The mean score was 4.3/5.0 (86%) with a standard deviation of 0.65.

Furthermore, the group also assessed the quality of each of the 50 sentences describing their personal scores for the colour energies. Below is a histogram of the results of this assessment.



Content Validity

This refers to the systematic determination of whether the content of a test measures the traits that it is designed to measure. The test developer attempts to build this type of validity into the test when it is constructed, through the selection of appropriate items. However, establishing proof that a test has content validity is only possible when what is being measured is a specific skill that independent experts agree an item or method can 100% verify e.g. an ear test involving the identification of musical notes could be proven to have content validity. Unfortunately for personality tests such as the Insights Discovery evaluator, there is no such agreement amongst experts on what constitutes good content validity for items describing psychological preferences. Consequently, content validity cannot be used as an approach to demonstrating the validity of the evaluator and in this case is as useless as face validity in making a claim to true validity.

From a scientific perspective, face validity is of marginal importance in establishing the validity of an instrument. Consider the opinion of H.L. Mencken below in demonstrating this point:

"The most common of all follies is to believe passionately in the palpably not true.

It is the chief occupation of mankind."

Overview of Construct Validity

As noted by Paul Kline in his book, *A Psychometric Prime (1997), p. 36-37:*

> "Face validity is not a guide to true validity, concurrent validity is applicable only where there are benchmark measures for the variables, and predictive validity, although powerful, is only effective where clear criteria can be established. … Content validity…is suited only to fields with specified skills and knowledge. To obviate these problems as far as possible, Cronbach and Meehl (1955) developed an approach to test validation, a test known as construct validity.
>
> In establishing the construct validity of a test the first step involves the definition and delineation of the meaning of the test variable. A construct, in the sense of construct validity, is essentially a concept. Hence delineating the meaning of the test variable means clarifying the nature of the concept to be measured."

Construct validity is a generic name given to a class of multivariate statistical methods whose primary purpose is to define the underlying structure of a data set. The underlying structure of the data is defined by a set of dimensions known as *factors*.

Factor Analysis can be used either in an exploratory or a confirmatory purpose. In the results show below, Factor Analysis has been used in a confirmatory perspective in order to test the theoretical hypothesis underlying the distinction between the Insights colour preferences.

In general, researchers use a 'rule of thumb' that considers factor loadings greater than 0.30 as meeting the minimal level for significance (Hair & al., 1998). Here are some key results based on a sample of 7,159 evaluators completed in English in the UK and further samples for evaluators for French-Canadian, French people in France, Germans and the Dutch. Many more factor analyses are available in more detailed papers.

	UK - English			
	S 3.0 (UK) 02-07/2004 N=7'159			
	Average factor loadings			
	F1	F2	F3	F4
Green	-0.56	-0.07	-0.13	0.07
Yellow	0.06	-0.30	0.48	0.04
Blue	-0.09	0.57	-0.23	0.02
Red	0.59	-0.20	-0.06	-0.04

Figure Seventeen – UK English Evaluator - Factor Loadings Summary Table

The F1 factor of 0.59 loads strongly onto Fiery Red. It also loads negatively onto Earth Green at -0.56. This negative correlation supports the theoretical construct of the model that hypothesises that 'Fiery Red' and 'Earth Green' are polar opposites.

The F2 factor of .57 loads strongly onto Cool Blue. It also loads negatively onto Sunshine Yellow at -0.30. This negative correlation further supports the theoretical construct of the model that hypothesises that 'Cool Blue' and 'Sunshine Yellow' are polar opposites.

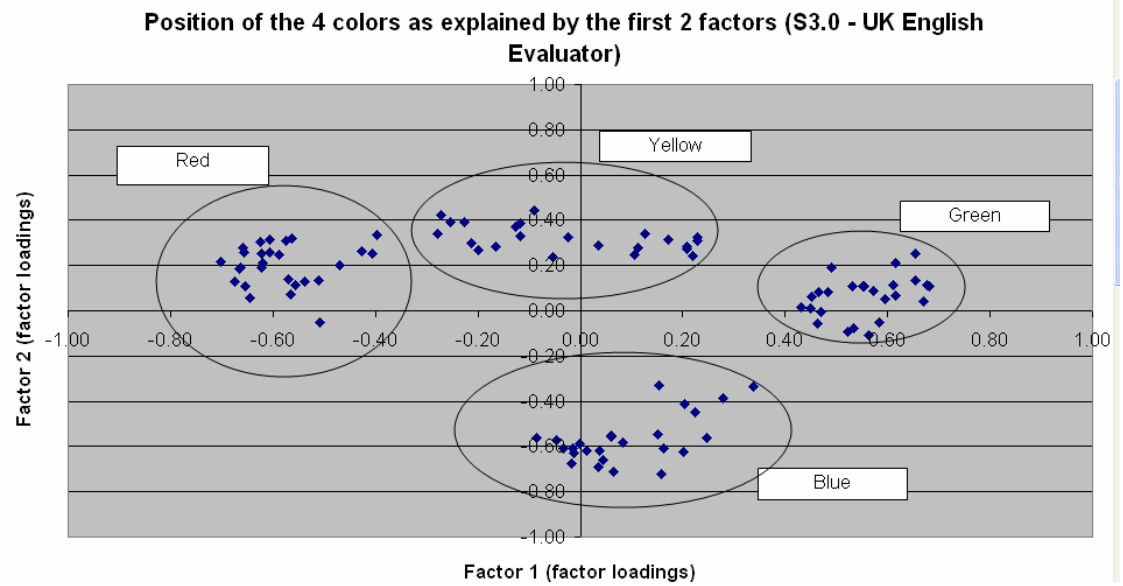


Figure Eighteen– UK English Evaluator – Graph of the 100 items (25 x 4 colours) plotted against the factors

The graph in figure eighteen shows the relationship of each of the 100 items (four colours multiplied by 25 frames) relates to the two factors. It provides further evidence of the bi-polar nature of the colour energies and the construct validity of the model.

Further results for other country's evaluators show a similar pattern.

	Germany - German					**Nethelands - Dutch**			
	S 1.0 (Ger) N=1'964					**S 1.2 (NL) Beta N=1'259**			
	Average factor loadings					Average factor loadings			
	F1	F2	F3	F4		F1	F2	F3	F4
Green	-0.25	-0.03	0.48	-0.12		-0.35	-0.08	-0.11	0.38
Yelow	0.07	-0.26	-0.07	0.48		-0.02	0.54	-0.18	-0.05
Blue	-0.12	0.50	0.02	-0.24		-0.11	-0.42	0.37	-0.06
Red	0.47	-0.14	-0.32	0.02		0.60	0.03	-0.06	-0.14

Figure Nineteen– German and Dutch Factor Loadings

	France - French				Canada - French			
	S 2.0 (FRE) N=1'570				S 2.1 (CAN) N=3'425			
	Average factor loadings				Average factor loadings			
	F1	F2	F3	F4	F1	F2	F3	F4
Green	-0.42	-0.08	-0.03	0.24	0.49	-0.15	-0.05	-0.03
Yellow	-0.03	0.50	-0.23	0.02	-0.06	0.51	-0.25	0.06
Blue	-0.06	-0.31	0.42	-0.06	0.09	-0.20	0.48	-0.09
Red	0.55	-0.06	-0.15	-0.07	-0.50	-0.11	-0.13	0.18

Figure Twenty– French French and Canadian French Factor Loadings

Technical Explanation of Construct Validity

The method used to determine the optimal number of factors is the latent root criterion or sum of squared loadings. The minimum number of factors to be extracted has been determined by the 'cumulative sum of squared loadings' that indicates the percentage of variance explained by the incremental factoring procedure. When the later factors do not significantly increase the total variance explained, it is debatable whether their inclusion adds value to the analysis.

The data on the Insights Discovery evaluator points towards an initial two-factor solution where the two first factors account for the bulk (27% to 42%) of the variance. One or two additional factors could be included in the factor solution, leading to a three or four factor solution, although these additional factors are only marginally significant (3%-4% each).

Looking more closely at the first two factors we can see that the colour opposite to the colour loading significantly in a particular factor, (for example, the colour opposite to 'blue' is 'yellow'), has a strong negative loading value. This may lead to the conclusion that the essence of the explanation of the four Insight color preferences is contained in the first two factors which explain the bulk of the variance. The presence of satisfactory loading values in further factors, which contribute only to a small increase in the variance explained, is an added value but not a pre-requisite to the validation of the psychometric tool.

As matter of comparison, other studies testing the validity of Jungian psychometric instruments, report the following results:

- Four distinct factors are correlated to the four MBTI constructs, accounting for 56% of the variance (Tzeng & al., 1989);
- Individual factors included in a four-factor solution are accounting for between 4% and 8% of the variance, and all factors are explaining 34% of the variance (Loomis & Singer, 1980);
- A two-factor solution was found in the MBTI, corresponding to the EI and the JP dimensions (Sipps & Alexander, 1987);
- Six distinct factors were found in the MBTI, of which four resembled the four Jungian scales (Sipps et al. (1985);

- A four-factor model vs. two competing five-factor models was found in the MBTI (Harvey et al. (1995).

The use of ipsative (forced-choice) scale in validity and reliability analysis

The narrow classical view about the use of Factor Analysis (for validity tests) and Cronbach-Alpha Analysis (for reliability tests), which are both derived from a correlation matrix, is that only 'interval' data types can be used. Ipsative (forced-choice) scales are based on 'ordinal' (i.e. ranked) data types and this 'forces' a correlation between items that artificially inflates the correlations in the correlation matrix i.e. the effect on a correlation matrix of items being scored using a forced-choice ordinal level scale is to attenuate the resulting correlations.

However, the Insights Discovery Evaluator uses a scale which is a hybrid between a forced choice scale and a Likert scale i.e. each item is given a score between 0 and 6 (a 7 point scale) and the forced choice is over 4 items. Although this Likert scale is still 'ordinal' Jaccard & Wan (1996). State that 'their use in statistical procedures which assume interval type data is commonplace'. Other authors also state that, 'the use of ordinal variables such as Likert scale with interval techniques is the norm in contemporary social science' (Labovitz, 1967, 1970; Kim, 1975; Binder, 1984).

Finally, the impact on the correlation matrix of a forced choice across 4 items, using a 7-point scale is likely to be significantly less than using a 'dichotomous' scale where a choice must be made one way or another. Here is an example from the 1942 version of the Gray-Wheelwright Jungian Type Survey (Wheelwright, 1964) that uses a 'dichotomous' scale:

At a party I

(a) like to talk

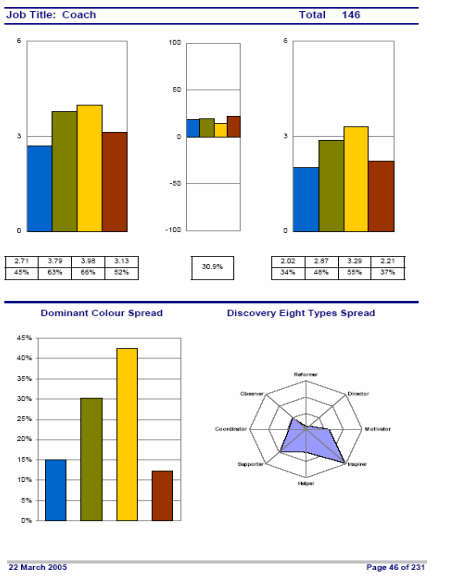
(b) like to listen

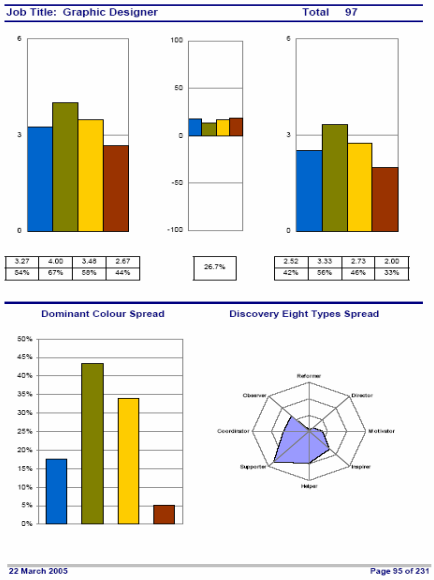
Choice (a) is a forced extraverted choice and (b) is a forced introverted choice. Numerous academic literature sources (Harvey et al.,1995; Tischler, 1994; Tzeng et al., 1989; Myers et al., 1998; Sipps & Alexander, 1987; Sipps et al., 1985) refer to the use of Factor Analysis as applied to 'dichotomous' measures e.g. as found in the MBTI (Myers-Brigg Test Instrument). These powerful statistical techniques, such as Factor Analysis and Cronbach's alpha, should be considered equally valid for the Insights forced-choice 7 point rating scale.

In summary, although at odds with the narrow classical view, there is sufficient evidence to support the valid use of these techniques on the Insights Discovery Evaluator data.

Criteria Validity

Criteria validity includes both predictive validity and concurrent validity. Concurrent validity studies are underway with the University of Westminster to statistically compare MBTI and the Insights Discovery Model. Predictive validity is evidenced by the data showing how aggregate data for different professions score differently across the colours. 35 professions have been analysed and 4 are presented here.





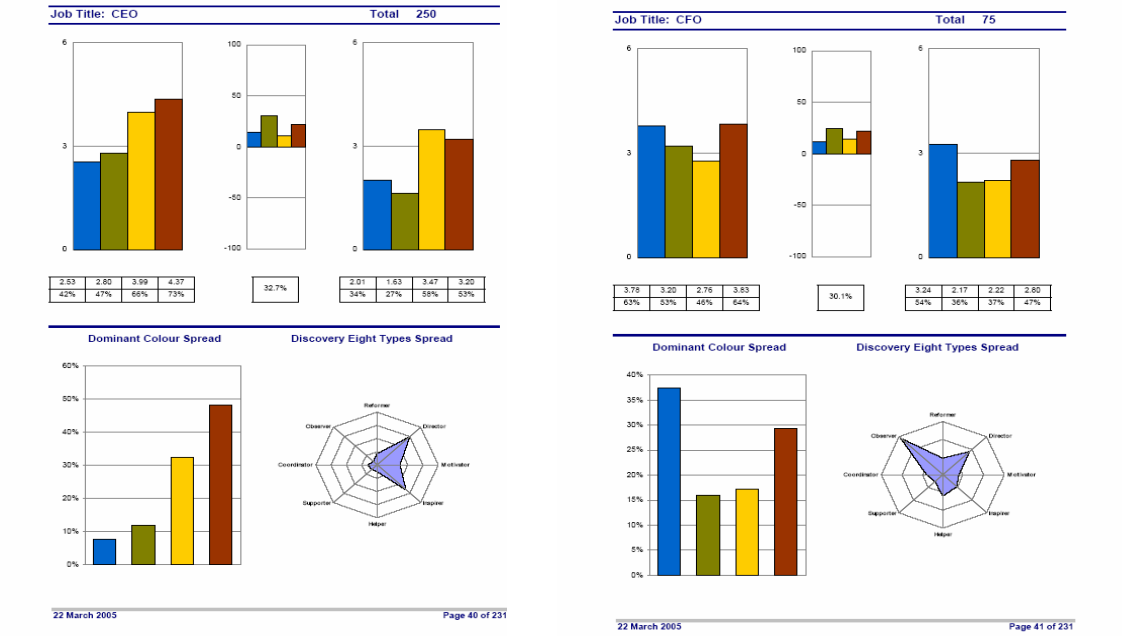


Figure Twenty One - A graphical view of some job description data supporting the argument for predictive validity

Below is a tabular display for the norms data that supports the argument for predictive validity.

The table shows one row for each different job description.

Population Segment	Sample Size	Average Colour Scores				Percentage of population with dominant colour energy			
		Blue	Green	Yellow	Red	Blue	Green	Yellow	Red
CEO	6,583	55%	58%	59%	54%	22%	25%	30%	22%
CFO	2,126	48%	55%	61%	60%	15%	21%	29%	35%
Coach									
Etc.									

Figure Twenty One - A tabular summary of the job description norms data

****** NOTE – TWEBTY ONE ABOVE NOT ACTUAL DATA …. Lindsay, please SEE POWERPOINT FOR THE ACTUAL DATA and input it******

While this data indicates that people in certain roles tend to have a *preference* for certain colour energies, it does not correlate or necessarily relate to *how well* they are doing they're job or *how capable* they are in fulfilling that role.

Conclusion

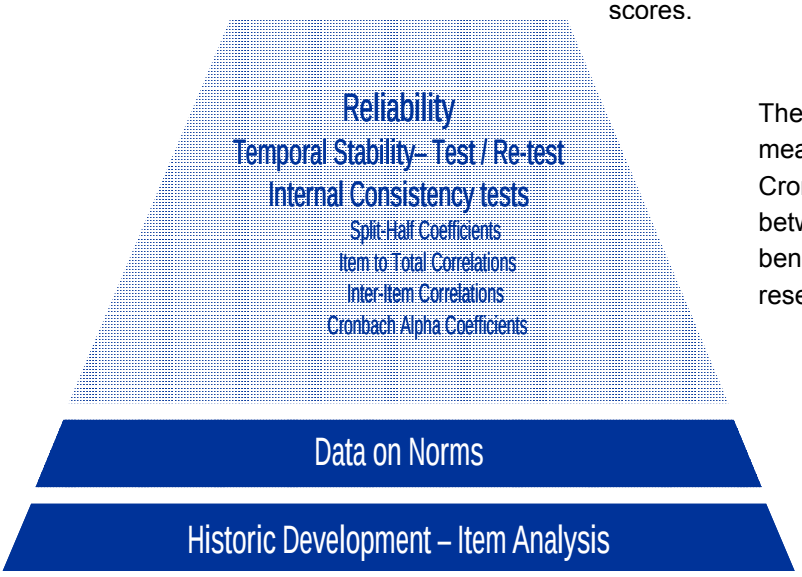


Figure One Repeated – Pyramid of Key Psychometric Statistics

There is strong evidence of construct validity as demonstrated by the factor analysis. Between two and four factors have been identified, with the first two factors typically explaining over 40% of the variance. The factor analysis data also provides evidence to support the Jungian interpretation that Cool Blue and Sunshine Yellow are polar opposites (as are Fiery Red and Earth Green), as evidenced by them loading both positively and negatively respectively onto the same factor.

Evidence of predictive validity is provided through an analysis of how the colour scores vary strongly by profession. e.g. Accountants tend to have higher Cool Blue scores and CEOs tend to have higher red scores.

There is strong evidence of the reliability of the measure of the four colours, as measured by the Cronbach Alpha and other statistics. Scores of between 0.91 and 0.93 compare favourably when benchmarked against other personality tests that research shows range between 0.7 and 0.9

Large samples of interesting norm data are available.

The development of the model through item analysis has been completed to a high standard.

This paper has explained how the Insights Discovery model has been developed through item analysis, supported by a large quantity of good quality data on norms. Building on this base, strong evidence of the model's reliability has been presented through the internal consistency tests and the test/re-test temporal stability data. The construct validity has been demonstrated through factor analysis and there is good predictive validity data by profession. These results all compare favorably with other Jungian based instruments that are held in high regard by psychometricians and meet the standards set out by both the American Psychological Association and the British Psychological Society. In summary, we have strong evidence to support the four colour measures calculated from the Insights Discovery model being both reliable and valid.

References for Reliability

APA (1999), 'Standards for educational and psychological testing', prepared by a joint committee comprised of the American Psychological Association, the American Educational Research Association and the National Council on Measurement in Education

Buley, J. (2000). Reliability, Validity and Correlation.

Churchill, G. A. (1979). A Paradigm for Developing Better Measures of Marketing Constructs. Journal of Marketing Research. 16 (February), 64-73.

Cronbach, L. J., & Meehl, P. E. (1955). Construct Validity in Psychological Tests. Psychological Bulletin, 52, 281 – 302

Dawis, R. V. (1987). Scale construction. Journal of Counselling Psychology, 34, 481-489

DeVellis, R. F. (1991). Scale Development: Theory and Applications, Sage Publications, Newbury Park, CA

Green, S.B., Lissit, R.W., Mulaik, S.A. (1977). Limitation of coefficient alpha as an index of test unidimentionality. Educational and Psychological Measurement. 37, 827-838

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., (1998). Multivariate Data Analysis, 5th ed, Prentice-Hall, Inc.

Henson, R. K., Kogan, L. R., & Vacha-Haase, T. (2001). A reliability generalization study of the Teacher Efficacy Scale and related instruments. Education and Psychological Measurement, 61, 404-420

Kline, P (1997) A Psychometrics Primer, Free Association Books, NYC

Kline, P (2000) Handbook of Psychological Testing, Routledge, 11 New Fetter Lane, London, EC4P 4EE

Nunnaly, J. C. (1972). Educational Measurement and Evaluation, 2nd ed. McGraw-Hill, New York, NY

Nunnaly, J. C. (1979). Psychometric Theory. McGraw-Hill, New York, NY

Peter, J.P. (1979). Reliability: A Review of Psychometric Basics and Recent Marketing Practices. Journal of Marketing Research, 16 (2), 6-17.

Robinson, J.P., Shaver, P.R. (1973). Measure of Psychological Attitudes. MI: Survey Research Centre Institute for Social Research, University of Michigan.

Robinson, J.P., Shaver, P.R., Wrightsman, L. S (1991). Criteria for Scale Selection and Evaluation. In 'Measure of Personality and Social Psychological Attitudes. Calif: Academic Press, San Diego.

Swailes, S., & McIntyre-Bhatty, T. (2002). The "Belbin" team role inventory: reinterpreting reliability estimates. Journal of Managerial Psychology, 17, 6, 529 – 536

Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20

Van Erkom Schurink, C. (August 2004). Measure Of Validity Through Factor Analysis, With A Focus On The Assessment Of The Various Versions Of The Insights 'Discovery Evaluator'. Working paper.

Van Erkom Schurink, C. (July 2004). The Issue Of Reliability And Validity In Psychometric Instruments. Working paper.

Wheelwright, J.B., Wheelwright, J.H., and Buehler, J.A. (1964) Jungian Type Survey (The Gray Wheelwright Test). San Francisco: Society of Jungian Analysts of Northern California

References for Validity

Cattell, R.B., (1966). The Scree Test for the Number of Factors. Multivariate Behavioural Research, 1 (April), 245-76.

Cliff, N., Hamburger, C.D., (1967). The Study of Sampling Error in Factor Analysis by Means of Artificial Experiments. Psychology Bulletin, 68, 430-45.

Gorsuch, R.L., (1983). Factor Analysis. Hillsdale, N.J.; Lawrence Erlbaum Associates.

Hair, J.F., Anderson, R.E., Tatham, R.L., Black, W.C., (1998). Multivariate Data Analysis, 5th ed, Prentice-Hall, Inc.

Kaiser,H.F., (1970). A Second Generation Little Jiffy. Psychometrika, 35, 401-15.

Kaiser,H.F., (1974). Little Jiffy, Mark IV. Educational and Psychology Measurements, 34, 111-17.

Loomis, M., & Singer, J. (1980). Testing the bipolarity assumption in Jung's typology, Journal of Analytical Psychology, 25, 4.

Remarczyk, M. (2005) MSc dissertation 'Face Validity of Insights Discovery' Available from the BPC, University of Westminster, 309 Regent Street, London, W1B 2UW, UK

Sipps, G. J., & Alexander, R. A. (1987). The multifactorial nature of extraversion-introversion In the Myers-Briggs Type Indicator and Eysenck personality inventory. Educational and Psychological Measurement, 47, 543-522.

Tzeng, O., Landis, D., & Chen, J. M. (1989). Measurement and utility of continuous unipolar rating for the Myers-Briggs Type Indicator. Journal of Personality Assessment, 53, 727-738.

Van Erkom Schurink, C. (2004). The Issue of Reliability and Validity in Psychometric Instruments. Working Paper. The Analytical Research Bureau, pp 40.

References concerning use of ipsative scales

Binder, A (1984). Restrictions on statistics imposed by method of measurement: Some reality, some myth. Journal of Criminal Justice, Vol. 12: 467-481.

Harvey, R. J., Murry, W. D., & Stamoulis, D. T. (1995). Unresolved Issues in the dimensionality of the Myers-Briggs Type Indicator. Educational and psychological Measurement, 55, 535-544

Jaccard, J, Wan, C. K. (1996). LISREL approaches to interaction effects in multiple regression. Thousand Oaks, CA: Sage Publications

Kim, J. O. (1975). "Multivariate analysis of ordinal variables." American Journal of Sociology, Vol. 81: 261-298.

Kim, J. O, Mueller, C. W. (1978). Factor Analysis: Statistical methods and practical issues. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences.

Labovitz, S. (1967). "Some observations on measurement and statistics." Social Forces, Vol. 46: 151-160Series, No. 14.

Labovitz, S. (1970). "The assignment of numbers to rank order categories." American Sociological Review, Vol. 35: 515-524.

Myers, I. B., McCalley, M. H., Quenk, N. I., & Hammer, A. L. (1998). MBTI Manual: A guide to the development and use of the Myers-Briggs Type Indicator. 3rd Ed. Palo Alto., CA: Consulting Psychologists Press.

Sipps, G. J., Alexander, R. A., & L. Friedt. (1985). Item Analysis of the Myers-Briggs Type Indicator. Educational and Psychological Measurement. 45(4), pp. 789-796

Sipps, G. J., & Alexander, R. A. (1987). The multifactorial nature of extraversion-introversion In the Myers-Briggs Type Indicator and Eysenck personality inventory. Educational and Psychological Measurement, 47, 543-522.

Tischler, L. (1994). The MBTI factor structure. Journal of Psychological Type, 31, 24-31

Tzeng, O., Landis, D., & Chen, J. M. (1989). Measurement and utility of continuous unipolar rating for the Myers-Briggs Type Indicator. Journal of Personality Assessment, 53, 727-738

References concerning Test-Retest Reliability scales

Capraro, M.R., & Capraro, M.M., (2000?). Myers Briggs Type Indicator Score Reliability Across Studies: A Meta-Analytic Reliability Generalization Study. *Working Paper (rcapraro@coe.tamu.edu).*

Carlson, J.G., (1985). Recent Assessment of the Myers-Briggs Type Indicator. *Journal of Personality Assessment, 49 (4), 356-365.*

Carlyn, M., (1977). An Assessment of the Myers-Briggs Type Indicator. *Journal of Personality Assessment, 41 (5), 461-466*

Harrell, T.H., & Lombardo, T.A., (1984). Validation of an Automated 16PF Administration Procedure. Journal of Personality Assessment, 49 (6), 638-642.

University of Westminster

bpc

Appendix A – Version 3.0 of the English Insights Discovery Preference Evaluator

INSIGHTS

INSIGHTS DISCOVERY PREFERENCE EVALUATOR

Introduction

This Evaluator forms the basis of your Insights Discovery profile. It is not a pass or fail test. It simply records your perception of your work preferences.

Instructions - Please read carefully

Find a time and place where you will not be interrupted.

1. Fill in the personal details section. Enter your name and the date on all three pages of the evaluator.
2. In each frame, read each word pair carefully. Select the word pair that MOST describes you in your work environment and circle **M** next to this.
3. From the remaining three word pairs, select the pair that LEAST describes you in your work environment and circle **L** next to this.
4. For each of the remaining two word pairs circle a weighting from the values 1, 2, 3, 4 and 5, where 1 represents 'not likely to describe me', and 5 represents 'very likely to describe me'. Please do NOT choose the same weighting twice. Select those weightings which you believe best represent the relative intensity of the description in your working personality.
5. Continue until all 25 frames have been completed. Please ensure every frame has been scored, and each of the four word pairs has been allocated an M, an L, or a value selected from 1, 2, 3, 4 or 5.

Guidance Notes

- Remember, this is NOT a test! There are no right or wrong answers.
- Respond to the Evaluator based on your perception of yourself. Do not discuss your choices with others.
- Choose your responses quite quickly, as your first impression is often best. As a guide, this Evaluator typically takes between 10-20 minutes to complete.
- If returning this evaluator by fax, only the pages containing the word pairs are required.
- Word pairs in different language versions of this evauator may not be compared directly. International versions are developed independently to ensure cultural differences are considered.

University of Westminster

bpc

INSIGHTS DISCOVERY PREFERENCE EVALUATOR Version S3.0 (UK)

Personal Details (please use BLOCK CAPITALS)

Date :____/____/____ (DD/MM/YY)
Title :_____ Male: ☐ Female: ☐
First Name :_____
Last Name :_____
Job Title :_____
Department :_____
Company :_____
Address :_____
:_____
:_____Postcode_____

Telephone :_____
Fax :_____
E-mail :_____
Date of Birth :____/____/____ (DD/MM/YY)
Staff No. :_____

Insights Use

0.	Sensitive and diplomatic	L 1 2 3 4 5 (M)
	Encouraging and valuing	L 1 2 3 (4) 5 M
	Precise and deliberate	L (1) 2 3 4 5 M
	Results-oriented and fast	(1) 1 2 3 4 5 M
1.	Composed and observing	L 1 2 3 4 5 M
	Diplomatic and calming	L 1 2 3 4 5 M
	Open and outgoing	L 1 2 3 4 5 M
	Active and controlling	L 1 2 3 4 5 M
2.	Amicable and quick	L 1 2 3 4 5 M
	Reliable and restrained	L 1 2 3 4 5 M
	Forceful and goal-oriented	L 1 2 3 4 5 M
	Methodical and logical	L 1 2 3 4 5 M
3.	Calm and even-tempered	L 1 2 3 4 5 M
	Determined and dominant	L 1 2 3 4 5 M
	Buoyant and light-hearted	L 1 2 3 4 5 M
	Exact and precise	L 1 2 3 4 5 M
4.	Confident and vigorous	L 1 2 3 4 5 M
	Orderly and concise	L 1 2 3 4 5 M
	Familiar and stable	L 1 2 3 4 5 M
	Talkative and genial	L 1 2 3 4 5 M
5.	Logical and clear	L 1 2 3 4 5 M
	Direct and challenging	L 1 2 3 4 5 M
	Loyal and accommodating	L 1 2 3 4 5 M
	Sociable and active	L 1 2 3 4 5 M

When you have completed the Evaluator, please ensure every frame has been allocated **ONE** 'M', **ONE** 'L' and **Two Different Values** selected from 1, 2, 3, 4, or 5.

University of Westminster

bpc

INSIGHTS DISCOVERY PREFERENCE EVALUATOR **Version S3.0 (UK)**

Name: _____ Date: _____/_____/_____ (DD/MM/YY)

6.	Relating and amenable	L 1 2 3 4 5 M
	Expressive and hopeful	L 1 2 3 4 5 M
	Powerful and assertive	L 1 2 3 4 5 M
	Thinking and self-contained	L 1 2 3 4 5 M
7.	Demonstrative and persuasive	L 1 2 3 4 5 M
	Questioning and reflective	L 1 2 3 4 5 M
	Initiating and self-confident	L 1 2 3 4 5 M
	Stable and concerned	L 1 2 3 4 5 M
8.	Resolute and confident	L 1 2 3 4 5 M
	Social and cheerful	L 1 2 3 4 5 M
	Faithful and helping	L 1 2 3 4 5 M
	Consistent and correct	L 1 2 3 4 5 M
9.	Sensitive and diplomatic	L 1 2 3 4 5 M
	Precise and deliberate	L 1 2 3 4 5 M
	Encouraging and valuing	L 1 2 3 4 5 M
	Results-oriented and fast	L 1 2 3 4 5 M
10.	In-charge and firm	L 1 2 3 4 5 M
	Reserved and cooperative	L 1 2 3 4 5 M
	Outgoing and gregarious	L 1 2 3 4 5 M
	Meticulous and detailed	L 1 2 3 4 5 M

11.	Team-focused and impulsive	L 1 2 3 4 5 M
	Accurate and rational	L 1 2 3 4 5 M
	Even-tempered and amiable	L 1 2 3 4 5 M
	Task-oriented and direct	L 1 2 3 4 5 M
12.	Analysing and painstaking	L 1 2 3 4 5 M
	Friendly and entertaining	L 1 2 3 4 5 M
	Competitive and robust	L 1 2 3 4 5 M
	Unassuming and responsive	L 1 2 3 4 5 M
13.	Constant and attentive	L 1 2 3 4 5 M
	Influencing and expressive	L 1 2 3 4 5 M
	Analytical and evaluating	L 1 2 3 4 5 M
	Bold and objective	L 1 2 3 4 5 M
14.	Strong-willed and purposeful	L 1 2 3 4 5 M
	Reasoned and particular	L 1 2 3 4 5 M
	Eager and engaging	L 1 2 3 4 5 M
	Concerned and sensitive	L 1 2 3 4 5 M
15.	Systematic and principled	L 1 2 3 4 5 M
	Fun-loving and popular	L 1 2 3 4 5 M
	Steadying and moderating	L 1 2 3 4 5 M
	Fast and reinforcing	L 1 2 3 4 5 M

Once you have allocated 'M' and 'L', and are weighting the two remaining word pairings on the scale 1, 2, 3, 4, or 5, please do not choose the same weighting twice.

University of Westminster

bpc

INSIGHTS DISCOVERY PREFERENCE EVALUATOR **Version S3.0 (UK)**

Name: _____ Date: _____/_____/_____ (DD/MM/YY)

16. Persuasive and animated	L	1	2	3	4	5	M
Decisive and immediate	L	1	2	3	4	5	M
Discreet and analytical	L	1	2	3	4	5	M
Tolerant and laid-back	L	1	2	3	4	5	M
17. Empathetic and patient	L	1	2	3	4	5	M
Contained and controlled	L	1	2	3	4	5	M
Task-focused and competitive	L	1	2	3	4	5	M
Discussing and spontaneous	L	1	2	3	4	5	M
18. Influential and informal	L	1	2	3	4	5	M
Considerate and empathetic	L	1	2	3	4	5	M
Impartial and evaluating	L	1	2	3	4	5	M
Challenging and determined	L	1	2	3	4	5	M
19. Prepared and systematic	L	1	2	3	4	5	M
Courageous and independent	L	1	2	3	4	5	M
Responsive and extraverted	L	1	2	3	4	5	M
Counselling and caring	L	1	2	3	4	5	M
20. Articulate and strong	L	1	2	3	4	5	M
Spontaneous and spirited	L	1	2	3	4	5	M
Studious and reasoned	L	1	2	3	4	5	M
Peaceful and harmonious	L	1	2	3	4	5	M

21. Organised and thoughtful	L	1	2	3	4	5	M
Patient and supportive	L	1	2	3	4	5	M
Strong and well-argued	L	1	2	3	4	5	M
Interacting and open	L	1	2	3	4	5	M
22. Objective and daring	L	1	2	3	4	5	M
Relaxed and peaceful	L	1	2	3	4	5	M
Factual and conventional	L	1	2	3	4	5	M
Lively and congenial	L	1	2	3	4	5	M
23. Animated and enthusiastic	L	1	2	3	4	5	M
Driving and realistic	L	1	2	3	4	5	M
Compassionate and considerate	L	1	2	3	4	5	M
Detailed and attentive	L	1	2	3	4	5	M
24. Supporting and steady	L	1	2	3	4	5	M
Independent and bold	L	1	2	3	4	5	M
Reflective and thorough	L	1	2	3	4	5	M
Good-mixer and lively	L	1	2	3	4	5	M
25. Cautious and accurate	L	1	2	3	4	5	M
Forthright and straightforward	L	1	2	3	4	5	M
Optimistic and upbeat	L	1	2	3	4	5	M
Accepting and loyal	L	1	2	3	4	5	M

When you have completed the Evaluator, please ensure every frame has been allocated **ONE** 'M', **ONE** 'L' and **Two Different Values** selected from 1, 2, 3, 4, or 5.

Insights Learning & Development, Tel: +44 (0)1382 908050 Fax: +44 (0)1382 908051