Selector Insight

User Manual and Technical Description

Contents

Introduction

Selector Insight is an online candidate screening and selection instrument developed to help identify the most suitable candidates for your work environment. Selector Insight is an affordable, easy to use selection tool designed by psychologists in accordance with strict scientific guidelines. Research consistently finds that the most effective method of selecting people for roles is ability and personality testing (Schmidt & Hunter, 1998). Selector Insight provides objective measures of candidate's abilities, personal styles, work preferences, job ideals and stress reaction.

Selector Insight is recommended for use in recruitment, team development, succession planning and for training and development purposes. The robust psychometric characteristics of the Insight scales coupled with the flexibility of online delivery make Insight an indispensable part of your recruitment and organisational development solution. This manual describes the rationale, recommended use, development, and psychometric characteristics of the Selector Insight.

Rationale

Increasingly organisations are recognizing the importance of recruiting the best candidate. Poor hiring decisions can be costly; in fact, research suggests that it costs an organisation approximately twice an individual's salary to replace them. Unfortunately, however, the greatest cost to an organisation can be when employees who don't fit the organisation remain.

Poor recruitment can result in decreased performance, increased interpersonal conflict, absenteeism, work place stress, a lack of organisational commitment, and ultimately an increased risk of personal grievances.

This document is comprised of two parts:

Part I is the user manual for Selector Insight. It contains guidelines for how the questionnaire should be administered and describes the various sections of the questionnaire. Details of each report scale are provided along with report interpretation guidelines.

Part II contains the technical description of Selector Insight. It outlines Selector Insight's development procedure and provides evidence for its reliability and validity. It describes the development sample and examines gender differences.

Part I:

Selector Insight User Manual

Selector Insight Questionnaire

The Selector Insight assessment is web-based. This means that the candidate's computer must be connected to the Internet for the duration of the questionnaire. To protect from Internet connection or web browser failure, Selector Insight keeps track of the last page completed. If the connection to an assessment is lost for any reason, the candidate can simply log back on and continue from where they were up to.

The questionnaire begins with a Welcome screen that summarizes the subsequent five sections.

Personal Details

This section collects personal data that is used to help ensure the assessment is not inadvertently discriminating against any group. Aside from the candidate's name and gender, none of the data collected appears in the report, nor does it affect the candidate's results.

My Personal Style

This section contains 53 questions that investigate the candidate's interpersonal, working and coping styles. Questions are posed in terms of how strongly one agrees or disagrees with various statements.

My Ideal Job

This section contains 64 questions related to work preferences comprise the My Ideal Job section. Candidates are asked how desirable various job characteristics are to them.

Ability Measure

The Ability Measure contains 30 questions that investigate the candidate's verbal, numeric and logical reasoning ability.

My Stress reaction

This section contains 24 questions that measure the candidate's typical reactions when they are under pressure. The questions ask the candidate to indicate whether various symptoms or behaviours occur less, the same or more when they are feeling under stress.

Administration Guidelines

Ensure the candidate has a quiet working environment and will not be distracted by phones or other people. Inform the candidate of the following points:

- Selector Insight is a 1 hour assessment that will provide an objective indication of:
    - How you relate to others
    - How you go about your work
    - What is important to you in your job
    - Your reasoning abilities
    - Your reactions when put under stress

- It is important to be as honest as possible when answering questions.

- The questionnaire requires you to answer all questions.

- You will not be penalized for guessing where you are unsure of the correct answer during the abilities section.

- It is a good idea to have a pen and paper available to help work out problems.

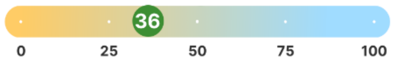
Selector Insight Report

The dimensions across which candidates are profiled have been selected for their relevance to screening and selection in employment settings, and are based on current psychological theory and best practice in applied psychology. This section describes the scales across which Insight assesses a candidate. Traits or tendencies of those that score at the upper and lower end of each scale are provided.

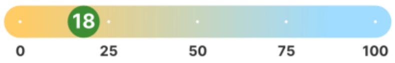
Ability assessment

The ability section reports overall reasoning aptitude and the component verbal, numerical and logical reasoning scores.

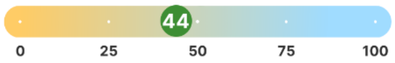
Overall Reasoning Aptitude

Below average		Above average

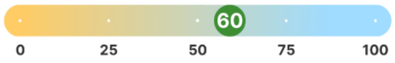
Verbal Reasoning

Below average		Above average

Numerical Reasoning

Below average		Above average

Logical Reasoning

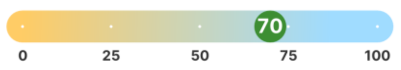
Below average		Above average

Personal styles

This section presents the candidate across personality dimensions that are consistently proven to predict job performance when overall job performance is the criterion, and when more specific criteria are being predicted. The results of this section are particularly useful in assessing likely team fit.

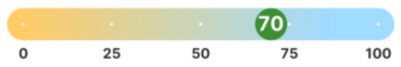
Competitiveness

This scale measures the need to compete or co-operate with others, how to meet goals and if a person measures themselves against others as an indicator of success.

Co-operative, prefers to work towards collective goals, flexible, willing to compromise		Competitive, values individual success, determined, goal driven

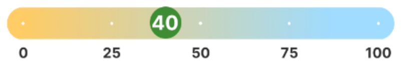
Extroversion

This scale measures the extent to which a person draws energy from interacting with others. Extroverted people tend to be outgoing, emotionally expressive, enjoy meeting and talking to people and are comfortable in social situations. Reserved people are quieter and reflective, and prefer more focused, smaller group interactions.

Reserved, quiet, prefers smaller social occasions, reflective, dislikes small talk, closed		Outgoing, extrovert, comfortable with large groups, seeks excitement, chatty, friendly, open

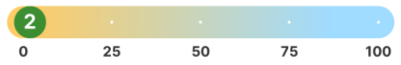
Openness to Ideas

This scale measures the interest in new ideas, approaches and experiences. High scoring people tend to be more curious, with an interest in concepts and theories and are more willing to debate ideas and opinions. Practical and pragmatic people tend to score lower, preferring to stick to the known or proven.

Practical, pragmatic, down-to-earth, prefers straightforward tasks, not academically inclined		Intellectually curious, likes to be challenged, philosophical, argumentative

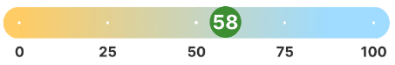
Orderliness

This scale measures the focus on order and structure. People with high scores regard reliability, responsibility, conscientiousness and constraint as being very important. If a person considers that spontaneity, quick reactions and variety are important; they will tend to score lower.

Disorganised, reactive, unconstrained, free-spirited, variety seeking, dynamic, undisciplined, untidy		Organised, planned, responsible, self-disciplined, finisher, detailed, process-focussed, stubborn

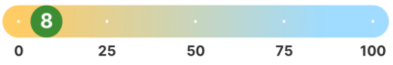
Self-Confidence

This scale measures the extent to which a person's sense of value, or worth, is based on their own views or on the opinion of others. The self-confidence scale embodies, self-esteem or belief in oneself. Those who rely on their own judgement tend to be less anxious and more at ease than those who refer to other people for their sense of self-worth.

Self-conscious, anxious, worrying, feelings easily hurt, sensitive to criticism		Confident, self-affirming, realistic appreciation of strengths and weaknesses

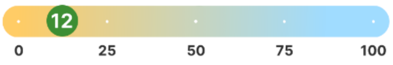
Teamwork

This scale measures the need to work together towards common goals, such as work targets, or to focus on individual goals. People with low scores like to operate independently, make their own decisions, and set their own directions. High scores indicate someone with a more collective approach, an active listener who is supportive of team members.

Independent, prefers to work alone, likes solo activities, self-contained		Sociable, team-oriented, collaborative, enjoys working towards shared goals

Tolerance

This scale measures whether the emphasis is placed on having tasks completed or if the people in their relationships should have a greater emphasis. People with high scores generally accept others as they are and try to maintain an even, patient manner in difficult or tense situations and avoid becoming angry or upset. Low scores indicate that the tasks have a greater importance to the person, they are more focused on what needs to happen, and they may be intolerant of interruptions at times.

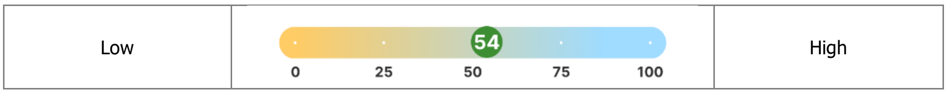
Direct, blunt, task-focussed, action oriented, tense, easily annoyed, intolerant		Patient, tolerant, people-focussed, slow to anger, easy-going, avoids upsetting others, avoids conflict

Ideal work environment

This section presents the candidate’s work preferences. Work preferences are important because people are more likely to excel in their work if they find it enjoyable. The results of this section can be used to assess the similarity between an individual’s preferred work environment and the work environment of the position for which they are applying. As for the Personal Styles section, this section reports scales in rank order.

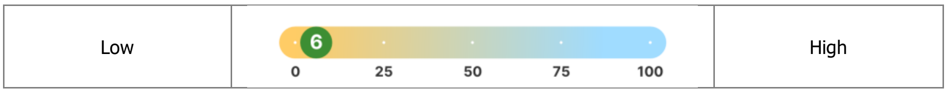
Autonomy

The Autonomy scale assesses the importance of having supervision and the ability to directly influence the nature of the work.



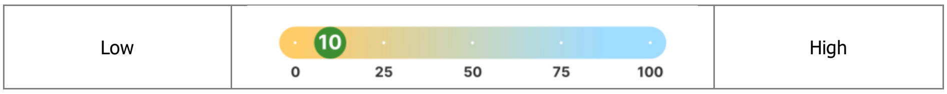
Complexity

The Complexity scale measures the importance of doing work that is either challenging and complex, or routine and straight-forward.



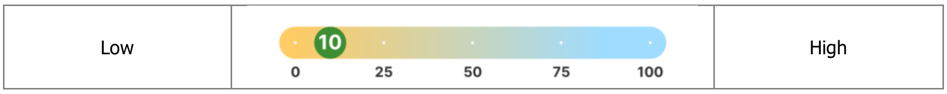
Interaction

The Interaction scale assesses the importance of how regularly interaction occurs with others in the work environment.



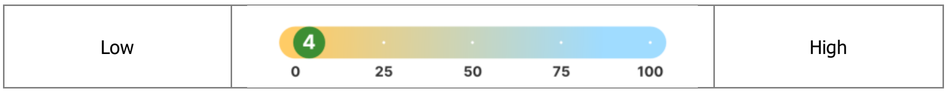
Physical

The Physical scale assesses the importance of the working environment, either outside or inside, and the level of physical work or exploratory activities involved as part of the role.



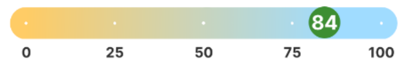
Predictability

The Predictability scale measures importance of stability, supportiveness and organisation in the workplace, and the value of security to the person.



Pressure

The Pressure scale assesses the importance of effort and commitment to a person, and how regularly the work will stretch and challenge them.

Low		High

Job ideals

This section presents the candidate’s responses to the My Ideal Job section of the questionnaire in two different ways.

First, it presents the candidate’s raw responses. Job characteristics are grouped by what the candidate finds ‘absolutely essential’ in a work role, right down to those that are ‘undesirable’. This presentation serves as a very convenient checklist for mapping what the candidate prefers with what a job role can offer.

Secondly, the job characteristics are presented in terms of how important they are to the candidate in comparison to others who have completed the questionnaire. This view can be used to temper the results of the previous view. For example, most candidates will answer positively to the ‘offers good pay’ characteristic. It is interesting to note however, whether compared to others, this characteristic is particularly important.

Ability scale summary

This section re-displays the Ability Measure scales for ease of comparison.

Behavioural scale summary

This section recasts each of the Personal Styles and Work Preference scales in a fixed order, making candidate comparison possible at a glance.

Stress reactions summary

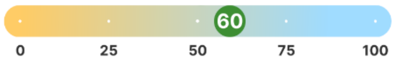
This section shows, at a glance, the candidate’s overall reaction to stress in comparison to the general population normative group.

Stress reactions details

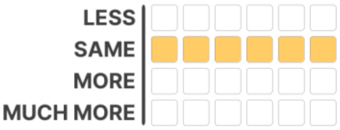
This section expands upon the Stress reaction summary to provide a detailed description of both the candidate’s reaction to stress and the individual scales that make up the overall stress reaction scale.

Stress reaction

Stress reaction refers to the overall pattern of stress reactions to stressful events. It is a summary of scores from the stress reaction second-order scales.

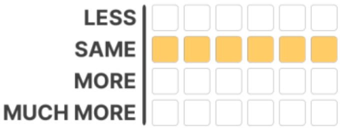
Less		More

Somatization



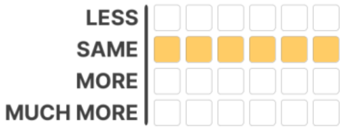
Somatization describes the physical experience of psychological symptoms, for example, the conversion of feelings of pressure from your environment into bodily dysfunction. These factors are linked to arousal of the autonomic nervous system, the part of the nervous system that controls involuntary body reactions.

Anxiety



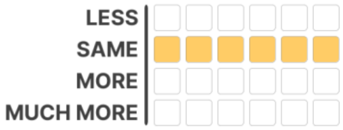
The Anxiety scale assesses the tendency to experience the cognitive aspects of stress, such as nervousness, tension and worry.

Distraction



The ability to concentrate when under pressure is a critical in many work environments, and it is well known that it is affected by workplace stress. Distraction measures the ability to focus on the task at hand without mental or thought blocks when under stress.

Withdrawal



Withdrawal measures the tendency to disengage from people and situations when events or situations become stressful.

Report Interpretation Guidelines

The following guidelines are presented to help increase the effectiveness of your interpretation of the Selector Insight report.

Ability Assessment

Consider the importance of various abilities for the role concerned. If a measured ability is not critical to the role then the score obtained is less important.

Verbal Reasoning

Verbal reasoning is a measure of the level of competency a person has with written language, spelling and meaning of words.

Important questions to consider:

- Is the ability to accurately convey meaning and express oneself in writing important to the role?
- How important are spelling and grammatical correctness to the job under consideration?
- Will the person be writing critical memos or letters to important clients?

Numerical Reasoning

Numerical reasoning is a measure of the level of competency a person has with numbers, numerical problems and the relationships between numbers.

Important questions to consider:

- Is the ability to work comfortably and easily with numbers important to the role?
- Will the person be dealing with critical accounts or data entry?
- What are the potential risks if numerical errors are made?
- If calculations are important to the role the application of a specific skill based test that mirrors the requirements of the role may be appropriate.

Logical Reasoning

Logical reasoning measures the ability to make inferences and solve complex problems given information from which a solution can be derived.

Important questions to consider:

- Is the ability to solve complex problems important to the position?
- Will the person be required to provide accurate and verifiable solutions to complex problems?

Personal styles and Ideal work environment

Selector Insight reports the behavioural scales in a consistent order, to allow ease of candidate comparison.

Determine the traits or success factors that are critical to a role. This should occur through job analysis. If a job description exists, this is likely to be the basis by which the critical success and fail factors are identified. Consider also those characteristics that may be undesirable in a role or result in a higher risk of failure.

Consider the candidate's work preferences. The extremes will be the most pertinent. Look to confirm a potential 'fit' or to identify a mismatch between candidate and role. Consider whether the candidate will indeed receive the Job Ideals indicated in the report that will motivate them to perform.

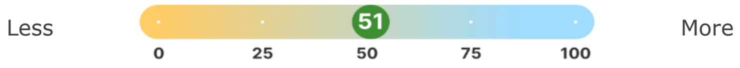
Utilize Selector Insight as a tool to provide direction for further investigation. Use interviews and reference checks to clarify or confirm any possible concerns and situational examples to probe for behaviour in the workplace.

Stress reactions

There is one overall *Stress reaction* scale and four second-order scales in the *Stress reaction* section of the report.

Stress reaction

The stress reaction scale is a report of the likelihood of experiencing anxiety, somatization, distraction and withdrawal, relative to others, when under stress.



Score	Interpretation Guideline
Less	Reports less physical and/or psychological stress than others. Maintains 'sameness' during stressful situations. May indicate low self-awareness, social desirability or 'hardiness'.
Average	Reactions to stress overall are in keeping with most other people.
High	Reports more physical and/or psychological stress than others. May be an indicator of counter-productive work behaviors. May indicate high self-awareness and honesty.

Contributors to stress reaction

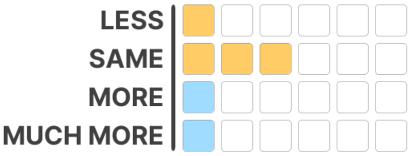
The four factors that contribute to stress reactions are:

- Anxiety *(the cognitive aspects of stress, such as nervousness, tension and worry when under stress).*
- Somatization *(the physical experience of psychological symptoms).*
- Distraction *(the loss of focus and memory when under stress).*
- Withdrawal *(the the tendency to disengage from people and situations when under stress).*

Each of these factors is measured with six test items. Individuals are asked the level to which they experience symptoms when under stress. The answers range from ‘less’, ‘same’, ‘more’ or ‘much more’ compared with when they are not under stress.

A horizontal histogram is provided for each factor so interpreters can understand the frequency distribution of answers to each of the test items.

In the example below one question was answered with ‘less’, three with ‘same’, one with ‘more’ and one with ‘much more’. When a candidate answers ‘more’ or ‘much more’ the symptom is reported so it may be investigated further.



Review the four stress reaction factors and consider the predominant answer style to each, as well as outliers.

Answer style	Interpretation Guideline
Less	Reports experiencing a reduction in this state when under stress. May indicate low self-awareness, social desirability or ‘hardiness’.
Same	Reports no change in this state when under stress. Maintains ‘sameness’ during stressful situations. May indicate low self-awareness, social desirability or ‘hardiness’.
More	Reports a physical and/or psychological reaction when under stress. Nb. Most of the population will experience some minor reactions under stress. Investigate individual and organisational impacts.
Much More	Reports a significant physical and/or psychological reaction when under stress. Investigate individual and organisational impacts.

Other considerations when using Selector Insight

Take variability into account. All forms of psychological assessment are subject to variability depending on factors such as how the candidate is feeling on the day, the purpose of the assessment, their understanding of the items and so on. This is known as *error of measurement* and it applies to all types of human evaluation. Blood pressure is a good example. A person's blood pressure can vary form day-to-day and even hour-to-hour. Doctors are aware of this and allow for it when making a diagnosis. With a psychological assessment, all scores must be treated as general indicators only, not as absolute measures.

Always obtain independent information. Because assessment results are subject to error of measurement and the assessment only covers a small spectrum of possible human behaviours, assessment results should never be used on their own. It is essential that interview or reference checking be conducted to independently assess observed patterns.

Don't rely on old assessment results. Assessment results have a limited life. If more than six months has elapsed then a new assessment may be required.

Ensure compliance with the relevant legislation. It is imperative that all relevant human rights and employment legislation is complied with in the use of this instrument. If you have any doubts or queries regarding appropriate usage please contact Selector Limited.

Part II:

Selector Insight Technical Description

Introduction to Psychometric Concepts

Psychometric assessments have demonstrated validity in selection settings (Schmidt & Hunter 1998). One of the fundamental advantages that psychometric testing has over other selection methodologies is that it adds standardization to the selection procedure. This means that all those in the selection process are exposed to the same selection procedure, which will help ensure equity in the selection process.

The demonstrated ability of selection instruments to add to the effectiveness of the selection process, coupled with the ability of psychometric assessments to provide insights into traits not easily assessable through other techniques, provide a compelling argument for the use of psychometrics in selection settings.

To fully capitalize on the benefits that psychometric testing can offer, it is important that the tests being used have sound psychometric properties. In psychometrics, there are two principle criteria that determine whether a psychometric test is appropriate in a given setting. These are the reliability of the assessment and the validity of the assessment.

The following two sections introduce these key concepts before applying them to Insight. For further reading on this topic we recommend consulting a book such as Kline's *Psychometrics Primer (2000)*.

Theoretical Basis and Construction

Ability Assessment

Theoretical basis

Research consistently shows that measures of cognitive ability are among the strongest predictors of future job performance (Schmidt & Hunter, 1998). Selector Insight measures cognitive ability across the *Overall Reasoning* scale and its component *Verbal Reasoning*; *Numerical Reasoning*; and *Logical Reasoning* scales, for the purpose of employee screening in recruitment and selection and organisational development initiatives such as training and team building.

Development Methodology

Item Development
An item set was developed covering the three major aspects of general cognitive ability: verbal, numerical and logical reasoning.

Scale Development
The item set was administered to 503 participants. Responses were then analyzed using principle component factor analysis. The decision to use principle components was made on the basis of a number of considerations.

While the mean of 5.01 for the first factor emerging (logical reasoning) from the orthogonal rotation of principle component analysis was the lowest, the second (numerical reasoning) emerged with a higher mean than the third (verbal reasoning), suggesting that the structure is not solely the result of item difficulty.

The final arbiter of whether or not to use a principle components solution was whether or not the factor structure made sense. In light of the factor structure that emerged, we are confident of real world meaningfulness. Verbal, numerical and logical reasoning are routinely found in the factor analysis of ability; furthermore the solution reflected our hypothesized scale structure.

Three scales emerged from factor analysis of the ability data:

1. *Verbal Reasoning* measures verbal fluency and word knowledge.
2. *Numerical Reasoning* measures number awareness and numerical analysis.
3. *Logical Reasoning* measures logical analysis and the ability to solve problems given relevant information.

Personal Styles

Theoretical basis

Research consistently finds that there are certain aspects of personality that predict future job performance. In particular, conscientiousness predicts across all occupations when overall job performance is the criterion (Barrick, Mount & Judge 2002, Hunter & Schmidt 1998). Emotional stability is also found to be a strong predictor when overall job performance is the criterion; however it is less stable as conscientiousness in terms of predicting more specific performance criteria.

The remaining factors of the big five model (extraversion, agreeableness and openness to experience) also predict performance, although these factors tend to do so for specific occupations and specific job performance criteria only. Taken together, the research described provides a compelling case for the use of personality testing in personnel selection and recruitment settings.

Development methodology

Item development

Items were developed based on the five factor model of personality and to cover aspects found to be critical to success and failure in the workplace from the experience of psychologist Keith McGregor.

Scale development

The items were administered to 503 participants and the results analyzed using principle components analysis. The factors that emerged after orthogonal rotation of seven factors (based on a scree test), and the dropping of ambiguous (double and triple loading) items and unrelated items (items loading less than 0.3) are as follows:

1. *Extroversion*
2. *Orderliness*
3. *Openness to Experience*
4. *Teamwork*
5. *Tolerance*
6. *Competitiveness*
7. *Self-confidence*

Two of the factors are not typically considered personality factors under the five-factor model: *Competitiveness* and *Teamwork*. While we can be confident from the factor structure that they are measuring constructs independent of the other scales, at this stage, they are supported only by face validity, and the extent to which they are artifacts of the process of factor analysis will ultimately be determined through predictive validity studies investigating the relationship between test scores and job performance criteria.

Work Preferences

Theoretical basis

Preference inventories identify job characteristics that a candidate finds desirable and undesirable. The reason for including a preferences inventory is based on the premise that people perform better in roles they enjoy.

Development methodology

Item development

Items were developed based on relevant psychological theory and the experience of psychologist Keith McGregor.

Scale development

The items were administered to 503 participants and the results analyzed using principle components analysis. The scree test suggested a six factor solution:

1. *Pressure*
2. *Autonomy*
3. *Interaction*
4. *Physical*
5. *Predictability*
6. *Complexity*

Stress reactions

Theoretical Basis

Costs to organisations and individuals due to stress, suggests that there is benefit to be gained through identification those more likely to respond to stressors, and to understand individuals stress response styles.
Research shows the role of individual differences in the stress process (Jex, 1998). The focus of research into the Insight stress reaction measure was to identify dimensions of current psychological stress that were likely to be representative of trait-based psychological reactions to occupational stressors.

Our hypotheses were based on the work of Jex (1998), Beehr & Bhagat (1985) and McGrath (1976).

Development Methodology

Item development

The item set for the measurement model was developed around the items and dimensions of the Brief Symptoms Inventory (BSI), which is a measure of psychological stress (Derogatis, 1977). The BSI is appropriate for measuring levels of distress in normal populations and is used extensively in occupational stress research (Jex, Bliese, Primeaux, 1977).

Dimensions were identified that we believed reflect genuine dispositional reactions to stress, rather than trait-based personality variables. Selector Insight measures an individual's disposition to experience specific strains in response to work stressors.

Scale development

The item set was administered to 3129 participants. Responses were then factor-analysed to reveal a correlated first order factor structure consisting of four factors related to an individual's stress reaction to stress and a single second order factor (implied by correlated factors from oblique rotation) measuring overall stress reaction.

Five scales emerged from the factor analysis of the data:

1. *Stress reaction*
2. *Somatization*
3. *Anxiety*
4. *Distraction*
5. *Withdrawal*

Reliability: Ability Measure

Test-Retest Reliability

If a test has high test-retest reliability, there is little chance that on a subsequent occasion a candidate will obtain a score that differs from their original score. It is crucial that test-retest reliability is high. If it is not high, either the test scales are unreliable, or the person has actually changed on the dimension in the period between the two testing occasions. In order to examine the test-retest reliability of a test, an assessment is made of the similarity between an individual's test scores over two occasions.

The test-retest reliability of a test is typically measured by a correlation coefficient, which varies between –1 and 1. A coefficient of –1 indicates a strong negative relationship between the two test scores, while a coefficient of 1 would indicate a perfect positive relationship between the scores. The closer the test-retest reliability is to 1 the stronger the relationship between two test scores. The benchmark set for test-retest reliability of Insight is 0.70. This would indicate strong test-retest reliability.

The data presented below represent the test-retest reliability of a subsection consisting of 165 participants in the original development sample over an interval of three months.

Ability assessment

The test-retest reliability coefficients for the Insight's Overall Reasoning and its sub-scales are strong, with just Numerical Reasoning falling below our stated goal of 0.70, and only by 0.01.

Scale	Reliability
Overall Reasoning	0.82
Verbal	0.70
Numerical	0.69
Logical	0.70

n = 165, all correlations significant at $p < 0.01$

Table 1: Test-retest reliability of the Ability Measure

Internal Consistency Reliability

An alternative indicator of the stability of a scale is split-half reliability. This splits the test up into two equivalent halves and assesses the relation between the two halves. The most common measure of split-half reliability is Kuder-Richardson KR-20. This can be proven mathematically to be the mean of all the possible split half reliabilities of a given test. Kuder Richardson KR-20 deals with the reliability of right-wrong response items, such as in the case of ability tests. A generalized formula altered to deal with multiple choice personality questionnaire items is alpha.

Because it is the average of all possible split-half combinations it is referred to as a measure of internal consistency of the test. The stronger the positive relationship among scale items and between items and the scale, the higher the internal consistency of the test, and the closer the coefficient alpha is to the test-retest reliability coefficient. Alpha can be thought of as a ratio of true variance to error variance. An alpha that is too high indicates that there could be redundancy in the scale; however an alpha that is too low indicates that the items in the scale are not measuring the same trait. Accordingly, an alpha level of between 0.7 and 0.9 is the standard that we aimed to achieve.

The standard error of measurement

Essential to evaluating the appropriateness of any statistical test is the concept of the standard error of measurement (SEM). This is a band that is placed around the score an individual obtains, and indicates that due to the non-perfect reliability of a scale, an individual's score may actually fall either side of the observed score. The smaller the standard error of measurement of a scale the more confident we can be of the accuracy of the measurement. The standard error of measurement is provided for all Insight scales.

The original internal consistency estimates and standard errors of measurement were estimated on a sample of *n* = 755 (*Selector Insight User Manual and Technical Description*, Selector Limited, 2003).

The sample for these analyses of the Ability Measure were *n* = 6889. Internal consistency estimates (coefficient α) for the total sample are higher than those previously reported (see Table 2). *Numerical Reasoning* is at the recommended criterion of 0.70. The *Verbal Reasoning* and *Logical Reasoning* sub-scales are well within acceptable ranges. The same patterns of result were evident when internal consistency for the overall test and its sub-scales were calculated for males and females. The SEMs are favourable.

Scale	# Items	Mean	Std Dev	Alpha	SEM
Total Sample					
Overall reasoning	30	19.023	5.337	.81	2.32
Verbal	10	6.50	2.27	.65	1.31
Numerical	10	7.61	2.16	.70	1.18
Logical	10	4.90	2.24	.62	1.38
Males					
Overall reasoning	30	19.82	5.13	.80	2.29
Verbal	10	6.80	2.22	.65	1.31
Numerical	10	8.00	1.88	.64	1.13
Logical	10	5.01	2.28	.64	1.37
Females					
Overall reasoning	30	18.30	5.42	.81	2.36
Verbal	10	6.23	2.28	.64	1.37
Numerical	10	7.26	2.31	.72	1.22
Logical	10	4.79	2.19	.61	1.37

Table 2: Internal consistency estimates for the overall Ability Measure and sub-scales

Descriptive Statistics

Table 3 indicates the difficulty of the Ability Measure items, most of the items were above the 0.60 level with question 1 being the easiest item. Some items are extremely difficult; however, these number only three, of which question 30 is the most difficult.

The average difficulty for the total test was 0.62 (SD = 0.17), *Numerical Reasoning* was 0.76 (SD = 0.09), *Verbal Reasoning* was 0.65 (SD = 0.12), and *Logical Reasoning* was 0.49 (SD = 0.19).

Table 3 shows the item-total correlations or, in other words, item discrimination. The average discrimination for the overall test was 0.32 (SD = 0.06). For the *Numerical Reasoning* sub-scale 0.35 (SD = 0.67), *Verbal Reasoning* 0.31 (SD = 0.52), and the *Logical Reasoning* 0.31 (SD = 0.60).

Item	***n***	**Mean**	**Std Dev**	**Item-total Correlation**
Q1	6889	0.92	0.27	0.24
Q2	6889	0.60	0.49	0.25
Q3	6889	0.69	0.46	0.44
Q4	6889	0.80	0.39	0.32
Q5	6889	0.78	0.41	0.32
Q6	6889	0.57	0.49	0.32
Q7	6889	0.68	0.46	0.44
Q8	6889	0.82	0.38	0.42
Q9	6889	0.90	0.29	0.31
Q10	6889	0.49	0.50	0.28
Q11	6889	0.77	0.42	0.33
Q12	6889	0.61	0.49	0.24
Q13	6889	0.63	0.48	0.31
Q14	6889	0.77	0.41	0.38
Q15	6889	0.80	0.39	0.31
Q16	6889	0.68	0.46	0.30
Q17	6889	0.61	0.48	0.32
Q18	6889	0.66	0.47	0.36
Q19	6889	0.70	0.45	0.30
Q20	6889	0.53	0.49	0.42
Q21	6889	0.64	0.47	0.28
Q22	6889	0.56	0.49	0.37
Q23	6889	0.70	0.45	0.36
Q24	6889	0.61	0.48	0.34
Q25	6889	0.70	0.45	0.35
Q26	6889	0.45	0.49	0.22
Q27	6889	0.47	0.49	0.36
Q28	6889	0.24	0.42	0.24
Q29	6889	0.26	0.44	0.33
Q30	6889	0.22	0.41	0.23

Table 3: Descriptive statistics

Validity: Ability Measure

Kline (2000) described validity as the extent to which a test measures what it purports to measure. Clearly validity is an important characteristic of psychometric tests. The question being asked when we investigate the validity of a test is whether or not the instrument is suitable for the use we intend. There are a number of approaches to assessing the validity of a psychometric test. The lowest level of validity is known as face validity – to answer this question we simply ask whether or not, at a surface level, the test appears appropriate for its intended use. This is insufficient justification for assessment of a test's appropriateness, and all good tests will have evidence of construct validity and criterion related validity.

Construct Validity

To demonstrate that a test has construct validity, we must first show that the test has interpretable factors or scales. By interpretable, we mean that the scales of a test are measuring separate constructs. Once we demonstrate that our scales can account for where one psychological construct ends, and a new one begins, we have the basis of construct validity. We demonstrate that we have interpretable factors in Insight through the statistical procedure of factor analysis. Factor structures demonstrate scale independence, and add to the evidence for the construct validity of assessments.

Confirmatory Factor Analyses (2006, *n* = 6889)

Confirmatory factor analysis (CFA) was used to assess the hypothesized factor structure of the Ability Measure. The underlying theoretical model for the measure was three first-order factors (*Numerical Reasoning*, *Verbal Reasoning*, and *Logical Reasoning*) supporting one second-order factor (*Overall Reasoning*). Figure 1 shows a diagrammatical representation of the Ability Measure. The main difference between traditional exploratory factor analysis and confirmatory factor analysis is that the latter allows for hypothesized models to be tested. Given the thirty items of the Ability Measure were written to theoretically measure three forms of reasoning, this meant that CFA was the most appropriate method to examine its factor structure.

Within the CFA framework one can judge how well the hypothesized model fits the actual data. In other words, the issue is how well the observed covariance matrix matches the specified model. There are many fit statistics used to evaluate fit, however the three most commonly used ones are: the root mean error of approximation (RMSEA; Steiger, 1990) where values of <0.05 are indicative of good model fit and values >0.05 and <0.08 are deemed acceptable model fit (Browne & Cudeck, 1993); the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), a relative fit index where values greater than 0.90 indicate adequate fit of the model to the data; and the Comparative Fit index (CFI; Bentler, 1990) where values greater than 0.90 are deemed to have adequate fit (Mulaik, James, Van Alstine, Bennett, Lind, & Stillwell, 1989). Goodness-of-fit for all invariance tests included the ▲χ^2 relative to the ▲df. The use of the Chi-square statistic was weighted less heavily due to its sensitivity to sample size. To assess the degree to which the various levels of invariance were attained the use of the ▲CFI was used (CFI $_{\text{constrained}}$ – CFI $_{\text{unconstrained}}$) with changes $\leq$ 0.01 indicating the null hypothesis for of invariance should not be rejected (Cheung & Rensvold, 2002).

The goodness-of-fit indices for the second-order model of the Ability Measure are presented in Table 5. The Chi-square was statistically significant indicating the model did not fit. Typically though this statistic is heavily influenced by sample size and therefore is more likely to result in models being rejected. The other goodness-of-fit statistics all suggested that the hypothesized model fitted the data, will all values being within the range of good fit.

An examination of the standardized beta weights (β) (Table 4) suggest that overall the three sub-scales each strongly identified the second-order dimension of reasoning (*Numerical Reasoning* β = 0.86, *Verbal Reasoning* β = 0.75, and *Logical Reasoning* β = 0.88. All factor loadings, except one item, item 30, were greater than 0.30 and therefore all contributed to the measurement of their latent variables.

Second Order Factor	First Order Factor	First Order Factor	First Order Factor	Factor Loadings λ
Numerical reasoning				0.86
Verbal reasoning				0.75
Logical reasoning				0.88
Q1	Numerical			0.31
Q3	Numerical			0.54
Q4	Numerical			0.39
Q7	Numerical			0.54
Q8	Numerical			0.54
Q11	Numerical			0.42
Q14	Numerical			0.46
Q15	Numerical			0.40
Q17	Numerical			0.39
Q19	Numerical			0.35
Q2		Verbal		0.33
Q5		Verbal		0.43
Q6		Verbal		0.42
Q9		Verbal		0.40
Q10		Verbal		0.36
Q12		Verbal		0.30
Q13		Verbal		0.38
Q16		Verbal		0.38
Q18		Verbal		0.46
Q20		Verbal		0.56
Q30			Logical	0.29
Q29			Logical	0.40
Q28			Logical	0.30
Q27			Logical	0.45
Q26			Logical	0.26
Q25			Logical	0.44
Q24			Logical	0.41
Q23			Logical	0.44
Q22			Logical	0.44
Q21			Logical	0.37

Table 4: Standardised factor loading for the second-order CFA model of the Ability Measure

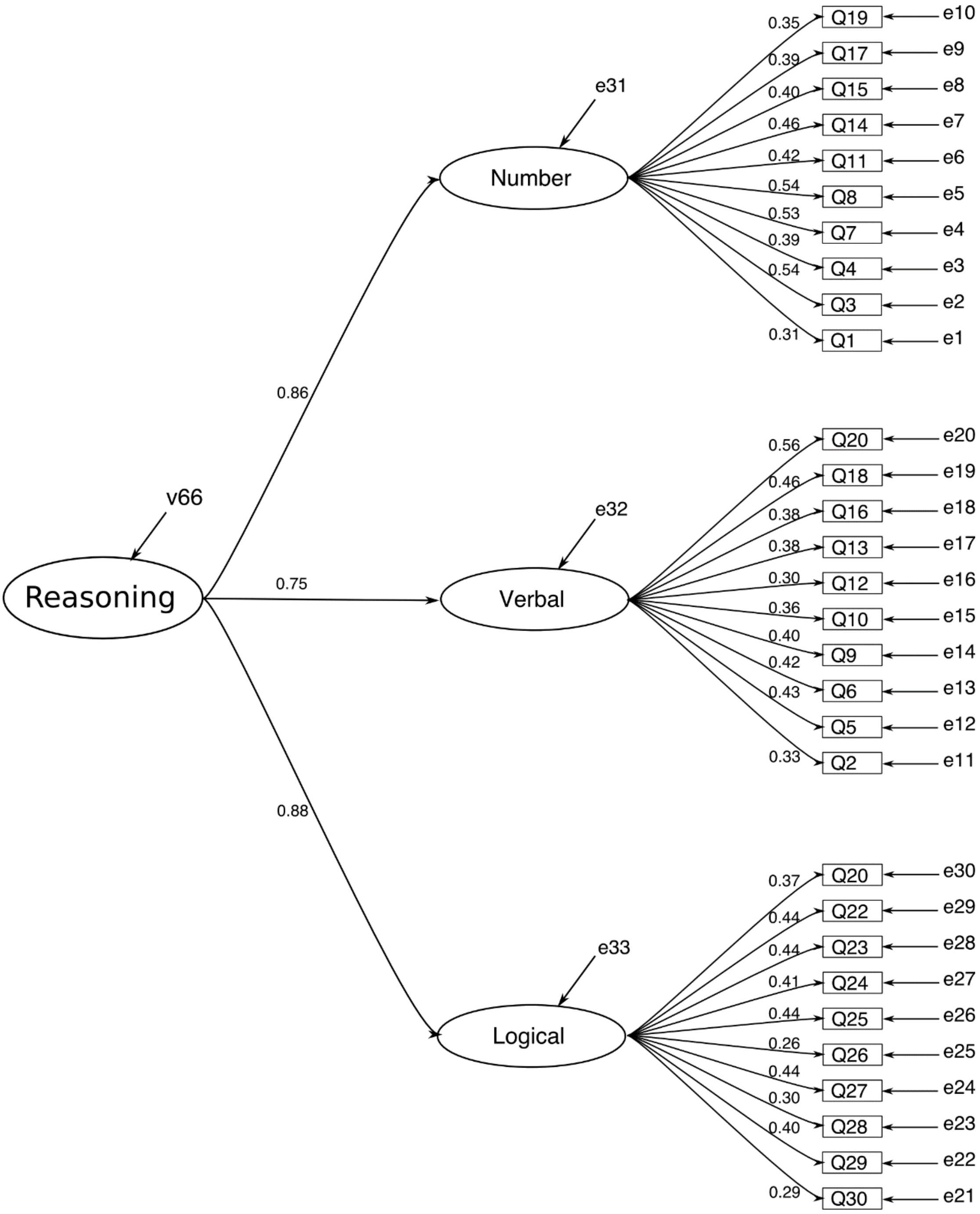


Figure 1

Model	df	x^2	P	TLI	CFI	RMSEA
3 factor second-order model	402	1635	0.000	0.94	0.94	0.021

Table 5: Fit Indices for the CFA for the Ability Measure

The gender invariance models specified were: *configural invariance*, in other words, the models were only specified to be the same between genders with no constraints added; *weak invariance* being equality constraints added to the factor loadings; and *strong invariance* with the added constraints of the item means being set to equality across the two groups.

The results for the gender invariance tests suggest that weak invariance was achieved as the ▲*CFI* for the difference between the weak and strong invariance models was great than -0.02. Generally the fit statistics for the weak invariance model were good (see Table 6).

	df	x^2	▲df	▲x^2	TLI	CFI	▲CFI	RMSEA
3 factor second-order model								
Independence	870	21341						
Configural	804	2116	64	19225**	0.93	0.93	0.00	0.015
Weak	833	2232	29	116**	0.93	0.93	0.00	0.016
Strong	863	2756	30	524**	0.91	0.91	-0.02	0.018

Note: ** = p<0.00

Table 6: Fit indices for the gender invariance of the Ability Measure

CFA is a sophisticated and robust methodology to use to confirm factor structures and the results from the above analyses are very encouraging in terms of adding validity evidence to the Ability Measure. Item factor loadings were moderate in identifying the three first-order factors, however, the loadings identifying the overall reasoning factor are very strong and suggest that verbal, numerical and logical reasoning all form part of an overall reasoning. The invariance models are also encouraging in that the measure managed to achieve weak invariance. While this sounds discouraging, one should remember that these types of models are very stringent in testing for equality across a range of parameters. The CFAs of the Ability Measure underpin the psychometric qualities of the measure and significantly enhance the factorial validity of this measure.

IRT

Factorial validity and internal consistency are part of the validity process and the more evidence one can accrue for a measure the greater the claims for construct validity. A powerful method for understanding more about the qualities of items is item response theory (IRT). IRT is particularly powerful in understanding how responses to questions relate to the underlying latent trait. IRT relates characteristics of items (item parameters) and characteristics of individuals (latent traits) to the probability of providing a particular response (Hambleton, Swaminthan & Rogers, 1991). As the Ability Measure is dichotomously scored then there are three models that can be applied to the data to test which is more appropriate:

i. A one-parameter logistic model which assumes all items have the same discrimination (*a*-parameter) and that items differ on difficulty (*b*-parameter).

ii. A two-parameter logistic model which assumes that items differ in both discrimination and difficulty.

iii. A three-parameter logistic model which assumes that items differ in both discrimination and difficulty and that guessing is present (*c*-parameter).

Once the best model was identified, the process included examining each item to determine its individual psychometric properties and to decide if it should be retained, remodelled or removed from the test. The results from the IRT analyses allowed for examination of the psychometric properties of items types to determine if certain response formats are impacting measurement precision.

Descriptive Statistics

Table 7 shows the item parameters for the thirty items of the Ability Measure under each of the three IRT models. Under the classical test model the mean item discrimination was *r* = 0.32 (SD = 0.06) and the mean item difficulty *p* = 0.62 (SD = 0.17). For the 1-PLM the mean item difficulty was *b* = -0.75 (SD = 1.07). The mean item discrimination and difficulty for 2-PLM were *a* = 0.94 (SD =0.23) and *b* = -0.71 (SD = 1.03) respectively. Under the 3-PLM the mean item discrimination was *a* = 0.62 (SD = 0.19), mean difficulty was *b* = -0.57 (SD = .99) and mean guessing *c* = 0.07 (SD = 0.11).

Item	*p*	*r*	*1-a*	*1-b*	*2-a*	*2-b*	*3-a*	*3-b*	*3-c*
Numerical									
Q1	0.92	0.24	0.920	-3.052	1.024	-2.820	0.590	-2.876	0.000
Q3	0.69	0.44	0.920	-1.036	1.410	-0.785	0.833	-0.751	0.020
Q4	0.80	0.32	0.920	-1.797	1.000	-1.696	0.583	-1.710	0.000
Q7	0.68	0.44	0.920	-0.983	1.366	-0.759	0.791	-0.762	0.000
Q8	0.82	0.42	0.920	-1.918	1.501	-1.395	0.871	-1.408	0.000
Q11	0.77	0.33	0.920	-1.538	0.992	-1.461	0.576	-1.477	0.000
Q14	0.77	0.38	0.920	-1.585	1.212	-1.313	0.702	-1.326	0.000
Q15	0.80	0.31	0.920	-1.789	0.937	-1.772	0.548	-1.783	0.000
Q17	0.61	0.32	0.920	-0.588	0.841	-0.635	0.494	-0.635	0.000
Q19	0.70	0.30	0.920	-1.114	0.881	-1.234	0.472	-1.245	0.000
Verbal									
Q2	0.60	0.25	0.920	-0.544	0.614	-0.762	0.358	-0.768	0.000
Q5	0.78	0.32	0.920	-1.646	0.933	-1.635	0.735	-0.688	0.371
Q6	0.57	0.32	0.920	-.0357	0.824	-0.395	0.515	-0.230	0.059
Q9	0.90	0.31	0.920	-2.746	1.216	-2.253	0.946	-1.224	0.509
Q10	0.49	0.28	0.920	0.018	0.691	.0015	0.405	0.015	0.000
Q12	0.61	0.24	0.920	-0.608	0.575	-0.900	0.375	-0.445	0.120
Q13	0.63	0.31	0.920	-0.708	0.791	-0.802	0.604	-0.143	0.224
Q16	0.68	0.30	0.920	-0.949	0.784	-1.080	0.456	-1.083	0.004
Q18	0.66	0.36	0.920	-0.873	0.990	-0.833	0.580	-0.817	0.009
Logical									
Q20	0.53	0.42	0.920	-0.143	1.172	-0.122	0.774	0.033	0.069
Q21	0.64	0.28	0.920	-0.723	0.712	-0.892	0.415	-0.899	0.000
Q22	0.56	0.37	0.920	-0.342	1.004	-0.325	0.653	-0.133	0.078
Q23	0.70	0.36	0.920	-1.110	1.024	-1.033	0.641	-0.824	0.096
Q24	0.61	0.34	0.920	-0.609	0.858	-0.648	0.514	-0.581	0.026
Q25	0.70	0.35	0.920	-1.081	0.980	-1.038	0.570	-1.046	0.000
Q26	0.45	0.22	0.920	0.219	0.523	0.336	0.391	0.848	0.135
Q27	0.47	0.36	0.920	0.157	0.992	0.147	0.779	0.436	0.125
Q28	0.24	0.24	0.920	1.459	0.734	1.741	0.791	1.630	0.100
Q29	0.26	0.33	0.920	1.289	1.065	1.164	0.815	1.179	0.050
Q30	0.22	0.23	0.920	1.578	0.740	1.872	1.209	1.556	0.122

Table 7: Descriptive statistics for CTT and IRT items parameter

Model Fit Evaluation

Equal Discrimination

The 1-PLM assumes that each item has the same discrimination value, and in many ways this is overly restrictive and not easily achievable. The 1-PLM, in an ideal world, would be the model of choice, but a real world perspective acknowledges that the 2-PLM or even the 3-PLM are more realistic models to apply as they allow for items to take on different discrimination values and guessing (3-PLM). A simple but effective approach to assess the degree to which the 1-PLM meets the assumption of equal discrimination is to examine the *point-biserial* correlations (r) to determine if there is any spread in their range. Examination of Table 7 suggests that the range in r was 0.20 indicating that equal discrimination was not evident. This evidence suggests that models that allow items to have unequal discrimination indices should be considered (i.e, either the 2 or the 3-PLM).

Item Parameter Invariance

A further test of model fit is to examine the invariance of the item parameters. Invariance is the cornerstone of IRT and can be tested by examining the relationship between item parameters across the subgroups that the test is intended to be used with (Hambleton, Swaminathan & Rogers, 1991). The data (n = 6889) was split in to two data sets: males (n = 3214) and females (n = 3618). Items were calibrated under each of the three IRT models for males and females separately. For all models the a, b, and c-parameters for males and females were then correlated and graphed to assess their degree of association. The degree of association between males and females for the various items parameters across the three IRT models are shown in Table 8. The results suggest that invariance was achieved within each model and therefore one of the central assumptions of IRT was well demonstrated.

	***b*-1-PLM**	***b*-2-PLM**	***a*.-2-PLM**	***b*-3-PLM**	***a*-3-PLM**	***c*-3-PLM**
R $_{male/females}$	0.97**	0.95**	0.80**	0.91**	0.82**	0.767**

** = $p < 0.00$

Table 8: Correlations between male and female sample for the 1, 2, and 3-PLM item parameters

Minimal Guessing

Guessing is particularly problematic where low ability examinees are faced with difficult questions (Hambleton, Swaminathan & Rogers, 1991). If guessing is not occurring within a group then the appropriateness of the 3-PLM is questionable as the estimation of the lower asymptote adds little useful information. To determine if guessing among low ability examinees was evident, those with overall scores at or below 13 ($\leq$40%) were selected (n = 855). A correlation was computed on their total score and their score on the five most difficult items (items 26-30). One would expect that low ability students would not get these items correct and thus the correlation between these two sets of scores should be zero or close to zero, which was in fact the case (r = 0.09, $p < 0.000$), suggesting that guessing was not evident within this group on these items, and thus the inclusion of the c-parameter adds little to explaining the item responses.

Summary of Model Fit Evaluation

While some statistical methods have been undertaken to ascertain which IRT model is most appropriate for the data it is, in the final analysis, based upon judgement of what misfits. From the above analyses one can rule out the 1-PLM model as there was varied discrimination among the point-biserial correlations. In effect this suggests that the stringent assumption of equal discrimination must be discounted and therefore either the 2-PLM or the 3-PLM model should be examined to determine its appropriateness.

The choice between the 2-PLM and the 3-PLM is difficult to make given that both showed evidence for model invariance. Given that the guessing was not particularly evident it is difficult to justify the estimation of the lower asymptote of the 3-PLM. That said, however, there is some useful information to be gleaned from this model and this will be discussed below. Given that there was little extra information to be had from the 3-PLM, as many items did not show any guessing, it is recommend that the 2-PLM is the most appropriate model to describe the data. The 2-PLM is a realistic model to use to measure ability as it is less restrictive than the 1-PLM. The results from these analyses show that under the 2-PLM that most of the items performed very well in terms of the psychometric properties.

Item and Test Information

Item information under the 2-PLM

One of the main advantages of IRT is that one can examine where items and the overall test do their most effective work. In other words, where is the greatest measurement precision? Under IRT, precision is indexed by item and test information. Item information is the reciprocal of the standard error (SE) at various ability levels, and generally items with high *a*-parameters tend to have higher measurement precision. When item information is high at certain ability levels the amount of error is low. Examination of the item parameters and ICCs suggested that 11 items had *a*-parameters greater than 1.00, which is typically considered high in ability tests. Of the remaining items, 12 had a-parameters $\geq$0.75 and $\leq$0.99, and 7 items had *a*-parameters <0.74. The high information items were spread throughout the scale with 3 being from the logical reasoning scale, 7 from the numerical scale, and two from the verbal scale.

Test information under the 2-PLM

One of the main advantages of ICCs is their additive properties. Summing the ICCs, results in the test information curve (TIC). Examination of Figure 2 shows that for the 30 items of the Ability Measure under the 2-PLM most of the information was around the -0.08 mark indicating that the test was towards the easy end of the continuum. Thus to make the test harder more items are needed in the upper end of the continuum to make the TIC more rectangular and therefore provide measurement precision at this part of the continuum.

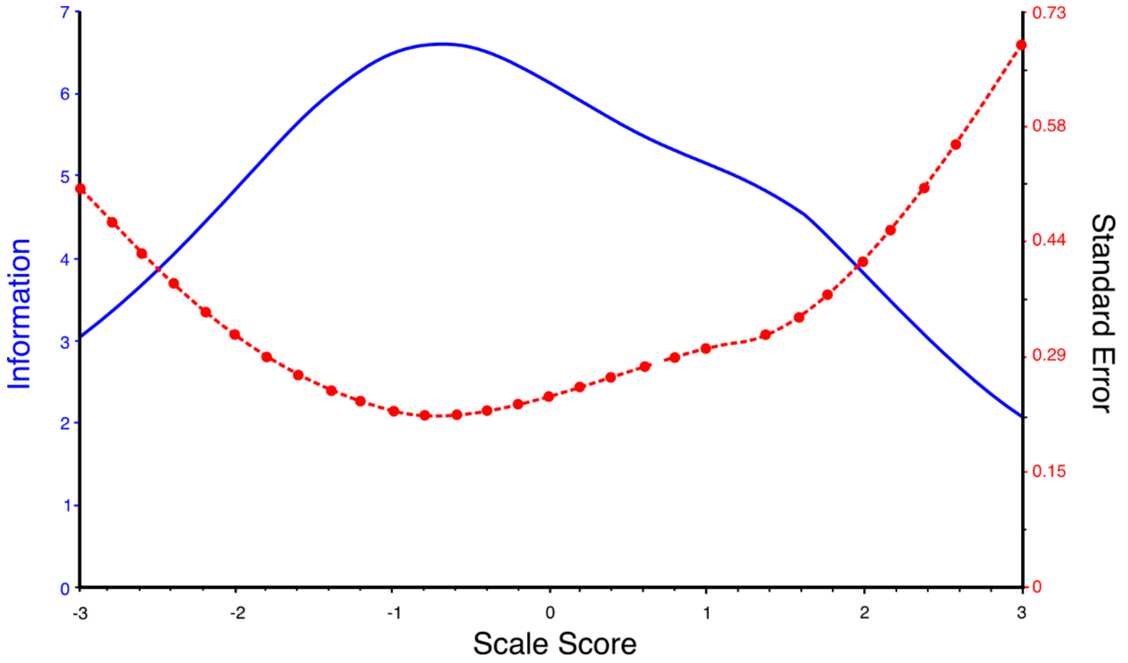


Figure 2: Test Information Curve for the 30 Item Ability Measure

Differential Item Functioning Analyses

An advantage of using IRT to analyses item responses is that one can then test the degree to which item responses are invariant across groups. The use of differential item functioning (DIF) provides an index of item bias. Item bias is typically considered to be related to secondary dimensions which are thought to be the cause/s of group differences. For example, take a math question which is worded in terms of a rugby context. If one were to match males and females on the basis of their math score for this item, one would hope that there would be no difference in the joint probability of each group answering the item correctly (no item bias). If, however, males had a significantly greater probability of getting the item correct compared to females, then this item would be biased in favour of males. Given that the question is based in rugby context one might conclude that this is primarily a male sport and that the bias in the item is a function of this context. The issue here relates to item dimensionality. One of the cornerstones of IRT is unidimensionality. In other words, one dimension accounts for an item response. DIF methods seek to identify such items, provide an estimation of the magnitude of the difference, and attempt to explain why differences occur.

For the following analyses multidimensional differential item functioning (MDIF) methods were employed (Shealy & Stout, 1993a, 1993b and Stout & Roussos, 1996). MDIF is based on the assumption that an item has one or more secondary dimension/s. For the following analyses SIBTEST (Stout & Roussos, 1996) was used. The data was divided into two groups: the reference group and the focal group. The focal group is typically the group the researcher believes to be disadvantaged by the item, whereas the reference group is the standard to which they are compared. Each sub-scale of the ability test was analysed separately and resulted in three runs of SIBTEST. SIBTEST provides a statistical estimate of the amount of DIF using a value called b_{UNI} which can be interpreted in a similar manner to the beta-weight in regression analyses using probability values.

Roussos and Stout (1996b, p. 220) proposed the following b_{UNI} values for classifying DIF on a single item: (a) negligible or A-level DIF is has an value of $b_{UNI} < 0.059$, (b) moderate or B-level DIF has an absolute value of $0.059 > b_{UNI} < 0.088$, and (c) large or C-level DIF with an absolute value of $b_{UNI} > 0.088$.

DIF Results

Table 9 shows the results for the MDIF analyses. Of the 30 items examined 15 were identified as having MDIF. One should, however, remember that it is the magnitude of the DIF that is important to consider when retaining/deleting items. Of the 15 DIF items only three items (items 11, 10 and 18) showed C level DIF which could be considered to be problematic. The remaining 12 DIF items exhibited a level of DIF that according to Roussos and Stout (1996) was negligible.

Four of the six DIF items (items 8, 11, 17 and 19) in the numerical section of the test favoured females. In other words, females when matched with a male of similar ability were more likely to get the items correct. Except for item 11, the remaining DIF items in this sub-scale showed negligible DIF and therefore do not present an issue. Item 11, on the other hand, certainly had a larger verbal component to its structure and this may be the cause of the DIF for this item.

For the verbal sub-scale 4 of the 5 (6, 9, 10, 13 and 18) items showing DIF favoured females. It would appear that there is something in the format of the questions for these items that may be leading females to answer correctly more than males. For this sub-scale, items 10 and 18 are clearly showing large levels of DIF and therefore their inclusion in the test should be closely scrutinized.

For the reasoning sub-scale the 4 DIF items (22, 24, 27 and 28) were split equally between males and females. Again the magnitude of the DIF is not of concern.

Item	Factor	Beta	p-value
Q1	Numerical	-0.009	0.183
Q3	Numerical	0.034*	0.007
Q4	Numerical	0.013	0.227
Q7	Numerical	-0.011	0.375
Q8	Numerical	-0.041*	0.000
Q11	Numerical	-0.068*	0.000
Q14	Numerical	0.026*	0.021
Q15	Numerical	-0.011	0.311
Q17	Numerical	-0.053*	0.000
Q19	Numerical	-0.027*	0.035
Q2	Verbal	-0.006	0.612
Q5	Verbal	-0.018	0.072
Q6	Verbal	0.051*	0.000
Q9	Verbal	-0.016*	0.027
Q10	Verbal	-0.076*	0.000
Q12	Verbal	-0.007	0.587
Q13	Verbal	-0.030*	0.016
Q16	Verbal	0.004	0.727
Q18	Verbal	0.084*	0.000
Q20	Verbal	0.008	0.493
Q21	Logical	-0.012	0.353
Q22	Logical	0.042*	0.001
Q23	Logical	0.008	0.458
Q24	Logical	-0.047*	0.000
Q25	Logical	-0.006	0.571
Q26	Logical	-0.009	0.512
Q27	Logical	0.050*	0.000
Q28	Logical	0.059*	0.000
Q29	Logical	0.008	0.509
Q30	Logical	-0.003	0.791

Table 9: MDIF results for the Ability Measure

Criterion Related Validity

Criterion related validity assesses the degree to which a test relates to appropriately selected criteria. These criteria may be other tests known to be effective measures of the construct being measured, or appropriately selected real world criteria.

We can ensure we have a valid assessment by ensuring that the test scales reflect the current state of theory, measure independent constructs, correlate with real world criteria and other measures known to reflect the construct, and appear relevant to inspection by laypeople.

Concurrent validity is the most common form of criterion related validity. It is more common than predictive validity because, as the name implies, measurements on both the predictor and the criterion are taken at the same time; making the need to wait for long periods of time before having criterion related data unnecessary.

Having demonstrated a basis for the construct validity of the Ability Measure, including the independence of the scales and the theoretical support for the proposed structure, it is important to demonstrate criterion related validity for the instrument. Accordingly, a study was undertaken to assess the level of correlation between the Ability Measure and educational achievement, a criterion known to correlate strongly with general cognitive ability.

The educational qualification of 503 participants was rated on a 6 point-scale, ranging from no formal qualifications (rating of 1) through to a doctoral qualification (rating 6). Of the 503 initial participants, 30 indicated unspecified professional or vocational qualifications, or other statements or certificates of achievement. Due to difficulty ascertaining the nature of these qualifications, these participants' data were removed from the sample. This left 470 participants, with educational qualifications rated on the scale presented in Table 10, below.

Educational Qualification	Rating
Doctorate or PhD	6
Masters or postgraduate (with or without honours)	5
Bachelors or postgraduate (with or without honours)	4
Trade certificate	3
Secondary school qualifications	2
No formal qualifications	1

Table 10: Classification system for educational achievement

The correlation between educational qualification and scores on the *Overall Ability Measure* scale and ability sub-scales was then calculated. Because of the multiple comparisons being made, Bonferoni adjustments were made to keep the overall significance level at a 0.05 level of significance. All correlations presented in the table below are significant at p = 0.0125.

Scale	Correlation	Corrected
Overall Reasoning Aptitude	0.35	0.39
Verbal Reasoning	0.29	0.38
Numerical Reasoning	0.18	0.22
Logical Reasoning	0.33	0.40

n = 470, all correlations significant at $p < 0.01$

Table 11: Correlations between the Ability Measure and educational achievement

The correlations in the corrected column of Table 11 have been corrected for unreliability. The correlation of 0.39 between *Overall Reasoning Aptitude* and educational qualification provides strong support for the predictive validity of the *Overall Reasoning* scale. The correlations between the sub-scales also provide sound evidence of the practical significance of scoring well in the Ability Measure. The benchmark for such coefficients to be of practical significance is 0.3 (Kline, 2000).

Conclusions

All the above analyses present a robust analysis of the Ability Measure of the Selector e-Profiler-II. The analyses used are a stringent test of the qualities of the items in test and the results from these analyses are very favourable.

From an IRT perspective the methods used to decide on model fit proposed by Hambleton, Swaminathan, & Rodgers (1991), suggested that the 2-PLM was the most appropriate. In saying this, one should be aware that information from the 3-PLM is instructive in terms of examining the content of items where guessing is taking place, albeit mild guessing. Typically items with *c*-parameters >0.20 are considered to be problematic items and are worthy of closer examination. Under the 3-PLM items 5, 9 and 13 all exceeded this criteria. While it is difficult to pinpoint why candidates are guessing, it is likely to be related item format, especially in the case of items 5 and 9. The wording of the item stems is confusing and this needs to be simplified (this is currently being tested as a reworded items). It is also worth noting that some guessing was apparent among many of the logical reasoning items, although none came close to the *c*-parameter of >0.020. Towards the end of the test some guessing was present. This is not surprising as these tended to be the hardest items in the test. One must therefore consider that there may be some order effect taking place. Thus examinees when faced with harder items at the end of the test are resorting to guessing.

The DIF analyses suggested that there were three items (items 10, 11 and 18) that need closer inspection to determine if they should be retained in the measure. On the basis of the magnitude of the DIF it is recommended these items be removed, remodelled and then retested. (New items are currently being trialled as replacements). Generally the levels of DIF found in the Ability Measure were small in magnitude and present no issue.

Reliability: Personal Styles & Work Preferences

Test-retest Reliability

The data presented below represent the test-retest reliability of a subsection consisting of 165 of the original participants in the development sample over an interval of three months.

Personal styles

The test-retest reliability data for the Personal Styles section is strong. With the exception of the *Self-Confidence* scale, all test-retest reliability coefficients are close to or in excess of 0.80.

Scale	Reliability
Extroversion	0.88
Orderliness	0.82
Openness to Exp.	0.81
Teamwork	0.80
Tolerance	0.79
Competitiveness	0.79
Self-Confidence	0.69

n = 165, all correlations significant at *p* < 0.01

Table 12: Test-retest reliability of Personal Styles

Work preferences

The test-retest reliability data of the Work Preferences scales is strong. All scales show test retest reliability coefficients of 0.70 or greater.

Scale	Reliability
Physical	0.82
Predictability	0.79
Pressure	0.79
Autonomy	0.70
Complexity	0.81
Interaction	0.84

n = 165, all correlations significant at *p* < 0.01

Table 13: Test-retest reliability of Work Preferences

Internal Consistency Reliability

An alternative indicator of the stability of a scale is split-half reliability. This splits the test up into two equivalent halves and assesses the relation between the two halves. The most common measure of split-half reliability is Kuder-Richardson KR-20. This can be proven mathematically to be the mean of all the possible split half reliabilities of a given test. Kuder Richardson KR-20 deals with the reliability of right-wrong response items, such as in the case of ability tests. A generalized formula altered to deal with multiple choice personality questionnaire items is alpha.

Because it is the average of all possible split-half combinations it is referred to as a measure of the internal consistency of the test. The stronger the positive relationship among scale items and between items and the scale, the higher the internal consistency of the test, and the closer the coefficient alpha is to the test-retest reliability coefficient. Alpha can be thought of as a ratio of true variance to error variance. An alpha that is too high indicates that there could be redundancy in the scale; however an alpha that is too low indicates that the items in the scale are not measuring the same trait. Accordingly, an alpha level of between 0.7 and 0.9 is the standard that we aimed to achieve.

The standard error of measurement

Essential to evaluating the appropriateness of any statistical test is the concept of the standard error of measurement. This is a band that is placed around the score an individual obtains, and indicates that due to the non-perfect reliability of a scale, an individual's score may actually fall either side of the observed score. The smaller the standard error of measurement of a scale the more confident we can be of the accuracy of the measurement. The standard error of measurement is provided for all Insight scales.

The internal consistency and standard error of measurement data presented below was calculated from a sample of 755 job applicants. The data was collected over the year from March 2002 to March 2003. The demographic data for these 755 people is presented in the Normative Base section of this manual.

Personal Styles

The internal consistency reliability of the Personal Styles section is strong. With the exception of the *Self-Confidence*, all reliabilities are over 0.7, indicating that the items within the scales themselves are measuring the same construct.

Scale	No Items	Mean	Std Dev	Alpha	Std Error
Extroversion	7	29.28	5.08	0.84	2.03
Orderliness	9	40.96	6.00	0.85	2.29
Openness to Experience	8	34.09	4.27	0.72	2.24
Teamwork	6	25.13	3.66	0.75	1.81
Tolerance	8	34.27	4.44	0.70	2.45
Competitiveness	8	29.37	4.89	0.72	2.58
Self-Confidence*	7	23.32	3.54	0.53	2.42

Table 14: Internal consistency reliability of Personal Styles

*Due to the lower internal consistency reliability of the Self-Confidence scale, it is recommended that the scale be used only as an indicator that must be backed up through exploration of information from another source, such as interview or referee check. Selector will continue to develop and refine this scale.

Work Preferences

The internal consistency reliability of the Work Preferences section is strong, providing evidence that the scales are measuring homogenous constructs.

Scale	No Items	Mean	Std dev	Alpha	Std error
Physical	5	14.74	5.31	0.83	2.20
Predictability	5	22.76	4.51	0.83	1.87
Pressure	5	16.75	4.34	0.75	2.18
Autonomy	5	23.71	2.96	0.66	1.73
Complexity	5	21.72	3.76	0.72	1.98
Interaction	5	22.39	3.49	0.68	1.98

Table15: Internal consistency reliability of Work Preferences

Validity: Personal Styles & Work Preferences

Kline (2000) described validity as the extent to which a test measures what it purports to measure. Clearly validity is an important characteristic of psychometric tests. The question being asked when we investigate the validity of a test is whether or not the instrument is suitable for the use we intend. There are a number of approaches to assessing the validity of a psychometric test. The lowest level of validity is known as face validity – to answer this question we simply ask whether or not, at a surface level, the test appears appropriate for its intended use. This is insufficient justification for assessment of a test's appropriateness, and all good tests will have evidence of construct validity and criterion related validity.

Construct Validity

To demonstrate that a test has construct validity, we must first show that the test has interpretable factors or scales. By interpretable, we mean that the scales of a test are measuring separate constructs. Once we demonstrate that our scales can account for where one psychological construct ends, and a new one begins, we have the basis of construct validity. We demonstrate that we have interpretable factors in Insight through the statistical procedure of factor analysis. Factor structures demonstrate scale independence, and add to the evidence for the construct validity of assessments.

Personal Styles

Scale	1	2	3	4	5	6	7
Extroversion 1	0.79						
Extroversion 2	-0.77						
Extroversion 3	0.77						
Extroversion 4	-0.73						
Extroversion 5	0.69						
Extroversion 6	-0.61						
Extroversion 7	0.57						
Orderliness 1		-0.73					
Orderliness 2		-0.71					
Orderliness 3		0.67					
Orderliness 4		0.67					
Orderliness 5		-0.66					
Orderliness 6		0.60					
Orderliness 7		0.59					
Orderliness 8		0.57					
Orderliness 9		-0.54					
Openness to Exp. 1			0.67				
Openness to Exp. 2			0.66				
Openness to Exp. 3			0.61				
Openness to Exp. 4			0.60				
Openness to Exp. 5			0.57				
Openness to Exp. 6			0.54				
Openness to Exp. 7			-0.53				
Openness to Exp. 8			0.46				
Teamwork 1				0.84			
Teamwork 2				0.74			

Scale	1	2	3	4	5	6	7
Teamwork 3				0.71			
Teamwork 4				0.66			
Teamwork 5				-0.64			
Teamwork 6				-0.58			
Tolerance 1					0.71		
Tolerance 2					0.67		
Tolerance 3					-0.66		
Tolerance 4					0.63		
Tolerance 5					0.60		
Tolerance 6					0.50		
Tolerance 7					0.41		
Tolerance 8					0.31		
Competitiveness 1						0.76	
Competitiveness 2						0.69	
Competitiveness 3						-0.66	
Competitiveness 4						-0.60	
Competitiveness 5						-0.48	
Competitiveness 6						0.43	0.30
Competitiveness 7						0.39	
Competitiveness 8						0.39	
Self-Confidence 1							0.57
Self-Confidence 2							0.55
Self-Confidence 3							0.55
Self-Confidence 4							0.49
Self-Confidence 5							0.47
Self-Confidence 6							0.45
Self-Confidence 7							0.44
% Variance explained	**7.85**	**7.71**	**6.76**	**6.43**	**6.15**	**5.91**	**5.02**

Table 16: Factor structure of Personal Styles

Work Preferences

Scale	1	2	3	4	5	6
Physical 1	0.82					
Physical 2	0.77					
Physical 3	0.74					
Physical 4	0.72					
Physical 5	0.70					
Predictability 1		0.76				
Predictability 2		0.73				
Predictability 3		0.72				
Predictability 4		0.72				
Predictability 5		0.67				
Pressure 1			0.74			
Pressure 2			0.73			
Pressure 3			0.68			
Pressure 4			0.67			
Pressure 5			0.63			
Autonomy 1				0.74		
Autonomy 2				0.74		
Autonomy 3				0.72		
Autonomy 4				0.59		
Autonomy 5				0.51		
Complexity 1					0.78	
Complexity 2					0.73	
Complexity 3					0.63	
Complexity 4					0.63	
Complexity 5					0.60	
Interaction 1						0.72
Interaction 2						0.65
Interaction 3						0.61
Interaction 4			0.32			0.59
Interaction 5						0.55
% Variance explained	**10.30**	**9.66**	**9.59**	**8.68**	**8.57**	**7.96**

Table 17: Factor structure of Work Preferences

Criterion Related Validity

In deciding what other tests to correlate the Insight scales with, we focused on ensuring that the scales used as baseline measures had construct level meaning. In much personality research, disagreement between scales scores on two tests is the result of a scale name that does not reflect the underlying dimensions being assessed. Indeed, this was much of the reason behind the early inconclusiveness of studies into the effectiveness of personality testing for the prediction of job performance in personnel selection situations.

Following the work of Hunter and Schmidt (1998) and more recently Barrick, Mount & Judge (2002) there is general agreement that the Big 5 personality dimensions can predict job performance. The strongest predictor of performance is found to be conscientiousness; followed by emotional stability. The remaining personality dimensions of extroversion, agreeableness and openness are found to predict well, but for more specific measures of job performance than overall performance (IPIP, 2001).

For this reason we selected the 50 item Big 5 factor markers from the International Personality Project, which can be found on the web at http://www.personality-project.org/perproj/online.html. This site contains the items that make up the marker factors used in the current study, their scale characteristics, and correlations with tests such as the NEO of Costa and McRae, one of the most widely known and replicated classification frameworks of personality in the world today. The table below contains our hypothesized direction of correlations that would exist between Insight's scales and the Big 5 factors.

Hypothesized Correlations

Insight Scales	Hypothesised Correlations with Big 5
Extroversion	Extroversion +
Orderliness	Conscientiousness +
Openness to Exp.	Openness +
Tolerance	Agreeableness +, Emotional Stability +
Self-Confidence	Emotional Stability +
Teamwork	Extroversion +
Competitiveness	Extroversion +
Physical	Openness -
Predictability	Neuroticism +
Pressure	Conscientiousness +
Autonomy	Conscientiousness + openness+
Complexity	Openness +
Interaction	Extroversion + Agreeableness +
Numerical	Openness +
Verbal	Openness +
Logical	Openness +
Overall Reasoning	Openness +

Table 18: Hypothesized correlations with Big 5 marker factors

Actual Correlations

Table 19, below, presents the correlations observed between the Insight scales and the Big 5 marker factors. Correlations approaching one, even prior to the standard correction for scale unreliability, indicate strong relationships between the two scale sets.

Insight Scales	Correlation	Corrected	Big 5 Factors
Extroversion	0.89	1.00	Extroversion
Orderliness	0.81	1.00	Conscientiousness
Openness to Exp.	0.64	0.78	Openness
Tolerance	0.28	0.35	Agreeableness
Tolerance	0.58	0.70	Emotional Stability
Self-Confidence	0.39	0.51	Emotional Stability
Teamwork	0.42	0.50	Extroversion
Competitiveness	0.31	0.37	Extroversion
Physical	*-0.01*	*-0.01*	*Openness*
Predictability	*-0.09*	*-0.11*	*Emotional Stability*
Pressure	*0.00*	*0.00*	*Conscientiousness*
Autonomy	*0.04*	*0.05*	*Conscientiousness*
Autonomy	0.44	0.57	Openness
Complexity	0.32	0.39	Openness
Interaction	0.55	0.64	Extroversion
Interaction	0.40	0.48	Agreeableness
Numerical	*0.10*	*0.13*	*Openness*
Verbal	0.30	0.39	Openness
Logical	*0.17*	*0.22*	*Openness*
Overall Reasoning	0.24	0.29	Openness

n = 165, all correlations significant at $p < 0.01$ unless italicized

Table 19: Observed correlations with Big 5 marker factors

The majority of the hypothesized correlations between Insight's scales and the Big 5 factor markers presented in Table 18 were found to exist in the hypothesized direction. Table 19, above, contains the actual and corrected correlations between Insight's scales and corresponding Big 5 factor markers.

Clearly there is a strong relationship between the Big 5 factor markers and the Insight scales. In particular, conscientiousness and extroversion stand out as being strongly related between the two scale sets. Research literature indicates conscientiousness and emotional stability are the most consistent predictors of subsequent job performance; followed by extroversion, openness and agreeableness with regard to more specific criteria. Table 19 above indicates that Insight's scales measure similar constructs to the five-factor model of personality, providing strong support for the use of Insight in selection and recruitment settings. More specifically, strong correlations were found to exist between extroversion, conscientiousness, openness and emotional stability.

The hypothesized negative correlation between the *Physical* scale and *Openness* was not observed, indicating these two scales are measuring different factors. Similarly, hypothesized correlations between the *Pressure* and *Autonomy* scales and the Big 5 factor of *Conscientiousness* were not observed in the study. Moderate correlations were found between the *Overall Reasoning* and reasoning sub-scales and *Openness*, a factor commonly found to correlate with intelligence.

Criterion Related Validity: In the Workplace

Expected Results

Given the known relationship between performance and cognitive ability (e.g. Schmidt & Hunter, 1998), we expected to observe strong correlations between the Insight Ability Measure and the ability sub-scales and both (a) managerial ratings of performance against corporate competencies and (b) managerial ratings of behavioural traits (on 5 point versions of the Insight scales), to the extent that these traits are job relevant.

Further, we expected to observe correlations above 0.2 between ability and job performance. The table below, taken from the O-net Testing and Assessment Guidelines, interprets the meaning of various value ranges of the validity coefficient, and highlights the rational for the target value of an r-value of 0.2.

Validity Coefficient	Value Interpretation
> 0.35	Very Beneficial
0.20 – 0.35	Likely to be Useful
0.11 – 0.20	Depends on Circumstances
< 0.11	Unlikely to be Useful

Table 20: O-Net Interpretations of validity coefficient size

It is important to keep in mind that significance is a function of the size of the effect and the sample size. The minimum significant correlation detectable for example, with a sample size of 65 (the size of the sample in this study) is 0.24 at $p < 0.05$, and 0.31 at $p < 0.01$. With this in mind we have drawn attention to significant relationships as well as substantive non-significant correlations that subject numbers did not permit the detection of correlation at a significant *p* level.

Reliability: Measure of Stress reaction

Internal Consistency Reliability

An alternative indicator of the stability of a scale is split-half reliability. This splits the test up into two equivalent halves and assesses the relationship between the two halves. The most common measure of split-half reliability is Kuder-Richardson KR-20. This can be proven mathematically to be the mean of all the possible split half reliabilities of a given test. Kuder-Richardson KR-20 deals with the reliability of right-wrong response items, such as in the case of ability tests. A generalized formula altered to deal with multiple choice personality questionnaire items is alpha.

Because it is the average of all possible split-half combinations it is referred to as a measure of internal consistency of the test. The stronger the positive relationship among scale items and between items and the scale, the higher the internal consistency of the test, and the closer the coefficient alpha is to the test-retest reliability coefficient. Alpha can be thought of as a ratio of true variance to error variance. An alpha that is too high indicates that there could be redundancy in the scale; however an alpha that is too low indicates that the items in the scale are not measuring the same trait. Accordingly, an alpha level of between 0.7 and 0.9 is the standard that we aimed to achieve, and as the tables indicate, this was achieved.

The standard error of measurement

Essential to evaluating the appropriateness of any statistical test is the concept of the standard error of measurement. This is a band that is placed around the score an individual obtains, and indicates that due to the non-perfect reliability of a scale, an individual's score may actually fall either side of the observed score. The smaller the standard error of measurement of a scale the more confident we can be of the accuracy of the measurement. The standard error of measurement is provided for all scales in the *Stress reactions* report.

Scale	Mean	Std Dev	N	Alpha	IIC
Stress reaction	56.10	11.37	148	0.92	0.35
Somatization	11.73	2.83	148	0.80	0.43
Anxiety	16.44	3.52	148	0.82	0.45
Distraction	14.05	4.05	148	0.88	0.57
Withdrawal	13.86	3.40	148	0.80	0.40

Table 21: Internal consistency reliability of the scales

Validity: Stress reaction Measure

Construct Validity

To demonstrate that a test has construct validity, we must first show that the test has interpretable factors or scales. By interpretable, we mean that the scales of a test are measuring separate constructs. Once we demonstrate that our scales can account for where one psychological construct ends, and a new one begins, we have the basis of construct validity. We demonstrate that we have interpretable factors in the Stress reaction Measure through the statistical procedure of factor analysis. Factor structures demonstrate scale independence, and add to the evidence for the construct validity of assessments.

Having established a basis for the construct validity of the Stress reaction Measure, we were interested in establishing its validity further by examining real world variables to which it is related. This section describes the relationship between the stress reaction measure and its sub-scales and other psychometric measures and job outcomes. In every case, the correlations reported are obtained from a validation sample of 148 volunteers from a call centre environment and employees in general management roles across a variety of organisations.

Other Psychometric measures: Personality

The first a priori hypothesis that we made was that the scales would have relatively low correlations with personality. If the Stress reaction Measure and its sub-scales have strong correlations with personality, it would add little incremental validity to selection decisions. Our hypothesis was that there would be small relationships with personality, as measured by five factor markers, other than for emotional stability. We know that people with low emotional stability describe their experiences as stressful regardless of their environmental influences.

	Extroversion	Agreeableness	Conscientiousness	Emotional Stability	Openness
Stress reaction	0.19	-0.01	0.14	0.39	0.09
Somatization	-0.13	0.01	-0.12	-0.19	0.04
Anxiety	-0.17	-0.10	-0.17	-0.38	-0.03
Withdrawal	-0.17	0.05	-0.10	-0.28	-0.06
Distraction	-0.15	0.06	-0.08	-0.38	-0.21

Table 22:Correlation with personality

To test this hypothesis we administered the Stress reaction Measure and Goldberg's five-factor marker set to the validation sample. The correlations in Table 22 indicate that our hypothesis is supported. In particular, the scale with the strongest relationship with *Stress reaction* and its sub-scales is emotional stability. The correlations between the emotional stability factor and *Stress reaction* and the stress reaction sub-scales are moderate. All are below 0.4 in absolute magnitude, indicating that less than 16% of the variance in these scales is shared.

Validity between the Stress reaction Measure and job outcomes is therefore more likely to be incremental, that is, over and above the predictive validity of personality measures. The correlations are also in the expected direction. People who are highly resilient are more emotionally stable. Conversely, people with high anxiety, distraction, somatization and withdrawal have lower emotional stability.

Among other correlations of note, Agreeableness and Openness had near zero correlations with *Stress reaction*, other than for the *Distraction* sub-scale of *Stress reaction* and Openness. *Stress reaction* and its sub-scales showed a modest positive correlation with Extroversion.

Other Psychometric Measures

	BSI Anxiety	BSI Obsessive	BSI Somatization
Stress reaction	-0.51	-0.45	-0.31
Somatization	0.30	0.18	0.32
Anxiety	0.45	0.35	0.20
Withdrawal	0.38	0.33	0.18
Distraction	0.51	0.56	0.33

Table 23: Correlation with BSI scales

Recall from the development description, we were interested in whether there was a stable trait-based response to stress that was different from an individual's current reaction to stress. The rationale being that just because someone has a propensity to respond to an environmental stimulus in a particular way does not mean that they will respond in that manner – it requires the presence of the stimulus to cause the response. Accordingly, we hypothesized that the Stress reaction Measure and its sub-scales would share a substantive portion of variance with scales on the Brief Symptoms Inventory (Derogatis and Melisaratos, 1983), but would not correlate strongly with these scales. The relationships shown in Table 23 above support this hypothesis.

Work Outcomes

	SIG Threat	SIG Pressure
Stress reaction	-0.33	-0.30
Somatization	0.20	0.15
Anxiety	0.29	0.27
Withdrawal	0.32	0.31
Distraction	0.27	0.25

Table 24: Correlation with Threat and Pressure

Having demonstrated the Stress reaction Measure's relationship to other psychometric variables, we assessed the relationship with work outcomes. First we tested the relationship between the Stress reaction Measure and current perceptions of the work environment, as measured by Hulin et al. (2003). This measure of general work stress has two sub-scales, the first measuring pressure, and the second measuring threat. The work published by Stanton and colleagues indicates the measure produces data with good reliability and validity. The correlations are all of sizeable magnitude and in the hypothesised direction. Highly resilient people perceive their workplaces as less threatening and less pressured.

Work Withdrawal and Job Withdrawal

	Work Withdrawal	Job Withdrawal
Stress reaction	-0.36	-0.23
Somatization	0.17	0.05
Anxiety	0.29	0.24
Withdrawal	0.33	0.26
Distraction	0.37	0.17

Table 25: Correlation with Work Withdrawal and Job Withdrawal

As a result of the confirmed relationship between stress reaction and perceptions of pressure and threat in the workplace, we hypothesised that low stress reaction workers would experience higher work withdrawal and job withdrawal. This scale, developed by Hanisch and Hulin (1991), measures one of the dimensions of organizational withdrawal. Work withdrawal is a broad construct that is defined by absenteeism, tardiness, and other behaviours that reflect employees' desires to avoid their work environment and work tasks. Each item describes a specific work withdrawal behaviour and asks the respondent how often they engaged in this behaviour on a 5-point scale from 0 (never) to 4 (many times). This was again confirmed, with correlations of sizeable magnitude and in the expected direction across the Stress reaction Measure and its sub-scales.

	OCB	Job Satisfaction
Stress reaction	0.22	0.17
Somatization	-0.25	-0.10
Anxiety	-0.19	-0.12
Withdrawal	-0.08	-0.15
Distraction	-0.23	-0.17

Table 26: Correlation with OCB and Job Satisfaction

Our job satisfaction measure consisted of newly developed scales from the Illinois Job Satisfaction Index (IJSI; Chernyshenko et al., 2003) that are designed to better capture some of the affective components of job attitudes. This scale is the second dimension of organizational withdrawal, representing employees' attempts to remove themselves from their job through turnover and/or early retirement (e.g., likelihood and desirability of quitting) (Hanisch & Hulin, 1991).

Organizational Citizenship Behaviours were assessed by a 12-item OCB measure. It is used to assess a variety of important behaviours that are generally not specified in job descriptions but are important for the successful functioning of an organization. This scale was adopted from Borman & Motowidlo (1997). As expected, sizable correlations with *Stress reaction* and its sub-scales were observed.

Testing the Measurement Model

The Stress reaction Measure developed as a result of our exploratory analysis in the previous section consists of 24 items that measure four first order factors related to stress and a single second order factor (implied by correlated factors from oblique rotation) measuring overall stress reaction. So, based on our earlier exploratory factor analyses of the responses of 2083 individuals, we have identified a correlated first order factor structure of an individual's stress reaction to stress, and a single second order factor we have called *Stress reaction*.

The four factors we identified as a result of the analyses are *Anxiety*, *Somatization*, *Distraction* and *Withdrawal*. For each item, respondents are presented with a four-item response scale on which they indicate the degree to which a symptom changes as a result of stress. The individual rates whether each stress reaction happens less when they are under stress, the same when they are under stress, more when they are under stress, or much more when they are under stress.

The confirmatory factor analysis model hypothesized a priori that (a) responses to the 24 items could be explained by four first order factors (anxiety, somatization, distraction and withdrawal) and a single second order factor (stress stress reaction); (b) that each item would have a non-zero loading on the first order factor it was intended to measure, and zero loadings on the other three first order factors; (c) error terms associated with each item would be uncorrelated; and (d) covariation among the four first order factors would be explained fully by their regression on the second order factor. This model is represented in Figure 4, below.

The response scale described above implies that items are measured on an ordinal level of measurement. When observed variables are on an ordinal, or combination of ordinal and interval scales, the categorical nature of the variables should be taken into account, in particular the SEM should not be estimated on the Pearson product-moment correlation matrix (Joreskog & Sorbom, 1996; Byrne, 1998). In the case of categorical data, they are based on the polychoric matrix and should be estimated using weighted least squares estimation. The estimation of model parameters using weighted least squares requires two matrices rather than one, the asymptotic covariance matrix and the polychoric matrix. We used Prelis to calculate the polychoric and asymptotic covariance matrices based on 1575 cases. In assessing the fit of the model we have very closely followed the recommendations of Byrne (1998).

The feasibility of the parameter estimates can be assessed by whether all items load on their appropriate factor, and in the appropriate direction. There are no parameters with unreasonable estimates. The standard errors are all acceptable, that is, not too small preventing definition of the test statistic or too large indicating the parameters cannot be determined. The test statistic provided by Lisrel is the *t*-statistic representing the parameter estimate divided by its standard error. It operates as a *z*-statistic in testing that the estimate is significantly different from zero. At the 0.05 level of significance the absolute value of the statistic must be greater than 1.96 before the hypothesis that the estimate is zero can be rejected. All loadings were significantly different from zero at this level of significance (test statistics not shown). Examination of the un-standardized solution has indicated all estimates are reasonable and statistically significant, and all standard errors appear appropriate.

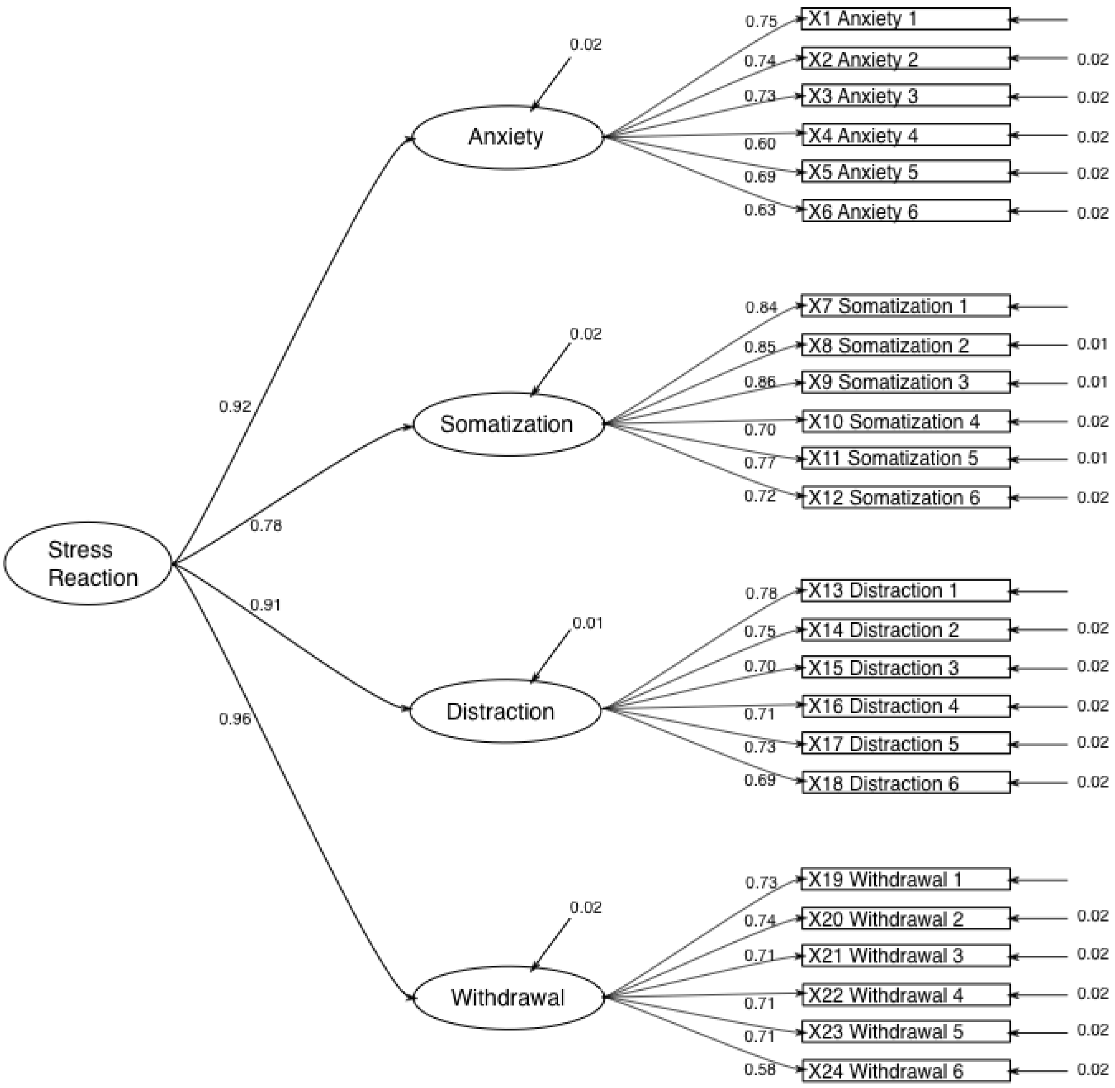


Figure 4

Measurement Model

The second step in assessing the fit is to examine the extent to which the measurement model is adequately represented by the observed measures. This is determined by the squared multiple correlations reported for each variable, which serve as reliability indicators of the extent to which each adequately measures its respective underlying construct (Bollen, 1989; Byrne, 1998). Examination of the squared multiple correlations for the 24 items (data not shown) indicated they were in general in excess of 0.5, with the minimum squared multiple correlation being 0.34 for item 13. Examination of the squared multiple correlations for the first order factors indicated they were 0.85, 0.61, 0.83, and 0.91 for *Anxiety*, *Somatization*, *Distraction* and *Withdrawal* respectively. These values provide relatively strong support for our hypothesized structure, indicating that a large proportion of the variance in the first order factors can be explained by our second order factor, *Reaction to Stress*.

Chi-square Goodness of Fit

The Minimum Fit Function Chi-square statistic tests the closeness of fit between the unrestricted sample covariance matrix and the sample covariance matrix; it tests the hypothesis that the difference between these matrices is zero. This statistic is equal to the sample size minus one times the minimum fit function, and is distributed as a central Chi-squared with degrees of freedom equal to ½(*p*)(*p*+1)-*t*, where *p* is the number of observed variables, and *t* is the number of parameters to be estimated (Bollen, 1989; Byrne, 1998).

The higher the probability associated with the Chi-square statistic the closer the fit between the hypothesized model under the null hypothesis and perfect fit, or alternatively, the *p* value is the probability of a Chi-square value exceeding the observed value if the null hypothesis was true (Byrne, 1998). The Chi-square statistic in this case was 1092.86 ($p = 0.0$), indicating the model is not adequate, indicating some causal misspecification. However there are problems that are widely known with regard to the use of the Chi-square test. In particular, the Chi-square based on the central distribution assumes that the model fits perfectly in the population. This is unrealistic in SEM research where postulated models can only ever approximate real world data (Macallum, 1998; Byrne, 1998). Further, the statistic is sensitive to sample size, and when is large it can lead to a significant Chi-square value even when the null hypothesis is tenable, which would lead us to reject the null hypothesis unnecessarily. Byrne noted this causes problems for SEM in which large samples are crucial to obtaining accurate parameter estimates. Chi-square tests are therefore unrealistic in most SEM empirical research (Byrne, 1998).

Byrne noted that when the hypothesis of no difference must be rejected the test is based on a non-central Chi-square distribution, and the resulting statistic is a non-centrality parameter (NCP), symbolized by lambda. The non-centrality parameter is a measure of the discrepancy between the observed and the estimated covariance matrices; the poorer the fit the higher the value of lambda. The Chi-square statistic is therefore a special case of lambda based on the central Chi-square distribution where lambda equals zero (Byrne, 1998). Our hypothesized model yielded a non-centrality parameter of 844.86, with 95% lower and upper bounds of 746.19 and 951.06, which is very wide.

Because of problems with the Chi-squared statistic, Jöreskog and Sörbom recommend it be interpreted as a measure of fit between the sample and the covariance matrices rather than a test statistic; a high Chi-square relative to degrees of freedom represents a poor fit while a small Chi-square relative to degrees of freedom represents a good fit. Measures other than the Chi-square/degrees of freedom ratio are now discussed, and are often termed subjective, practical or ad hoc indices of fit (Byrne, 1998).

Root Mean Square Error of Approximation (RMSEA)

The root mean square error assesses how well a model with unknown but optimally chosen parameter values would fit the population covariance matrix if it were available (Byrne, 1998). This is sensitive to the complexity of the model, with values of zero indicating perfect fit; less than 0.05 indicating good fit; and values as high as 0.08 representing reasonable errors of approximation. The RMSEA for the current model was 0.047, indicating good fit. The indication of good fit is supported by the narrow range of the 90% confidence interval, and in particular the upper bound being under the value of 0.05. The *p* value of close fit is greater than 0.98, and greater than 0.5 as suggested by Jöreskog and Sörbom (1996) and Byrne (1998). This evidence suggests the hypothesised data fits the model well.

Expected Cross Validation Index (ECVI).

The ECVI assesses the likelihood that the model will cross validate across similar sized samples from the same population, measuring the discrepancy between the fitted covariance matrix and the expected covariance matrix that would be obtained in another sample of equivalent size (Byrne, 1998). Byrne notes that application of the ECVI assumes a comparison of models whereby they are ranked by the magnitude of the ECVI (as the ECVI can take on any value) and the smallest ECVI statistic represents the model most likely to generalize. To evaluate the fit of the model using ECVI we compare its value with that of the saturated model and the independence model. The independence or null model (complete independence of all variables in the model, the most restricted) represents one end of a continuum and the saturated model represents the other end (where the number of parameters equals the number of variances and covariances of the observed variables, the least restrictive), and the hypothesized model is in between. The ECVI was actually lowest for the saturated model (0.38), indicating that if a new random sample was to be obtained the saturated model would cross validate better than our hypothesized model, although the value of the ECVI for the hypothesized model (0.76) was considerably lower than the null model (6.22), indicating it is more likely to generalize. The confidence intervals around the hypothesized model supported this interpretation (0.70, 0.83).

Chi-square Test for the Independence Model

We expect the Chi-square value for the independence model (with higher degrees of freedom due to the estimation of error terms only) to be much greater than the Chi-square value for the hypothesized model discussed earlier. This is the case for our model (1092.86 with 248 degrees of freedom versus 9744 with 276 degrees of freedom). Relative to the null model our hypothesized model is a substantially better fit to the data.

Aikaike's Information Criterion and Bozdogan's Consistent Aikaike Information Criteria

Similar to the ECVI criterion these two measures represent the likelihood that the model will generalize, and are interpreted in the same manner as the ECVI. That is, the model with the lowest value (independence, hypothesized and saturated) is the model most likely to generalize to new random samples. Similar to the results of the ECVI the AIC indicated that the model most likely to generalize was the saturated model (600), however the value for the hypothesized model indicated that it was more likely to generalize than the null model (1196.86 versus 9792.70). Conversely, the indication from the Bozdogan's CAIC indicated that the hypothesized model was the model most likely to generalize, followed by the saturated model and the independence model (1527.69, 2508.60, and 9945.39 respectively).

The Root Mean Square Residual

The standardized root mean square residual represents the average value across all standardized residuals. This value is standardized to overcome the fact that the un-standardized residual value is relative to the sizes of the observed variances and covariances. Interpreted in the metric of the correlation matrix, 1 represents poor fit and zero represents perfect fit. Byrne appears cautious in recommending below 0.05. The value observed in the current model of 0.13 means that the model explains the correlations among the observed and hypothesized correlation matrices with an average error of 0.13. Ideally this value would be lower.

The GFI and AGFI

The goodness of fit and adjusted goodness of fit indices represent absolute measures of fit, they compare the fit of the model with no model at all. The only difference between the adjusted and the goodness of fit indices is that the AGFI accounts for the degrees of freedom in the specified model. These values can be negative which would indicate that the hypothesized model fits worse than no model at all, which is clearly undesirable. The values for the current model of 0.98 and 0.97 for the goodness of fit and adjusted goodness of fit approach unity and indicate that the hypothesized structure fits the data better than no model at all.

The Parsimony Goodness of Fit Index

This indexes the goodness of fit of the hypothesized model while at the same time accounting for the parsimony of the model. The values are generally lower than the goodness of fit indices previously described, with values as low as 0.50 not unexpected even with non-significant Chi-squared values and goodness of fit indices at around 0.90. The PGFI for the current model of 0.81 can therefore be taken as consistent with previous fit statistics, and even suggestive of stronger fit than previously indicated.

Bentler and Bonnett’s Normed Fit index (NFI) is an incremental index of fit that compares the fit with the independence model rather than comparing with no model as with the previous goodness of fit indices. Similar to the previously described goodness of fit indices, the Comparative Fit Index can be viewed as version of the NFI that accounts for the fact that the NFI underestimates fit in small samples. These statistics range from zero to one with 0.90 indicating good fit. The values of NFI (0.89) and CFI (0.91) indicate good fit to the data. We do not present the Non-Normed Fit Index as Byrne argued that it is difficult to interpret because the values can exceed one.

The Incremental Index of Fit developed by Bollen was 0.91. As with the CFI, this was developed to address the susceptibility of the effects of NFI to sample size, and the similar value to the CFI is therefore unsurprising – both indicate a good fit of the hypothesized model to the data. The Relative Fit Index is algebraically equivalent to the CFI in most SEM applications, and ranges from zero to one. In this case the two values were not equivalent, yet at 0.88 it a similar value to the CFI of 0.91.

The Parsimony Normed Fit Index (PNFI) addresses the complexity of the model in the assessment of goodness of fit. The value for the current model (0.80) is within acceptable limits. Finally, the Critical N (CN) focuses directly on the adequacy of the model rather than the fit of the model. The CN estimates a sample size that would yield an adequate model fit for the Chi-square test. Our sample at over 1500 was closer to three and a half times the CN value of 437.01.

Summary of the Measurement Model

On the basis of the observed pattern of statistics indicating a good fit between our hypothesized model (based on our previous research) and the data, we are confident the structure adequately reflects theoretical and empirical considerations. We chose not to respecify the model on the basis of residual analysis, which would have taken us into an exploratory mode of analysis for specification searches. As the model is theoretically meaningful and supports our previous exploratory research on a large sample, we are confident with the factor structure of our hypothesized model, and chose not to risk an over specified model which would decrease the generalizability of the results.

Case Studies

Case Study One

Relationships between Insight and performance in a call centre environment.

Analysis indicated the level of difficulty of the Insight Ability Measure is suitable for call centre selection. In particular, evidence that Insight Ability Measures are of an appropriate level come from three sources.

(a) There are no floor or ceiling effects in the data for overall ability (nobody scored zero out of 30 and nobody scored 30 out of 30).

(b) The measures of central tendency are all very close to the centre of the possible range (i.e. 15 out of 30). There was a mean of 16.57, and a median of 17. While there are multiple modes, these are all presently above the mean, and we expect the mode to converge on the mean as the number of subjects who complete the test increases;

(c) Sub-scale analysis indicates that the means of numerical and verbal reasoning are both slightly above the centre of the possible range of scores, with only logical reasoning having a mean slightly below the middle of the range.

	n	Average	Median	Minimum	Maximum	Std dev
Total	56	16.57	17.00	4	28	5.57
Numerical	56	6.86	7.50	1	10	2.67
Verbal	56	5.82	6.00	1	10	2.10
Logical	56	3.89	4.00	0	9	2.01

Table 27: Descriptive statistics for ability scores of call centre staff

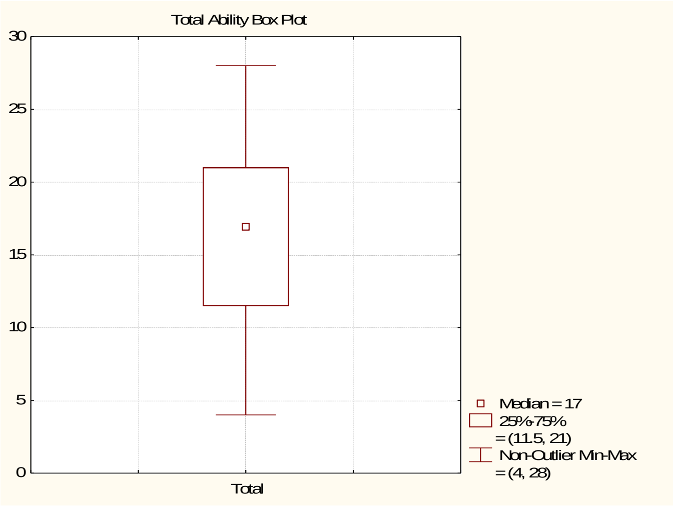


Figure 5: Box plot of total ability scores from a call centre environment

Relationships between call centre performance and Insight

Criterion	Best Predictor	*r*	*p*
Average Handling Time	Orderliness	-0.28	0.08
Call Coaching	Physical	-0.44	0.00
Communication	Predictability	-0.30	0.02
Customer Service	Predictability	-0.23	0.08
Detail Consciousness	Verbal	0.25	0.06
Flexibility	Work Pressure	0.29	0.03
Self Management	Work Pressure	0.20	0.13
Tardiness	Total Ability	0.29	0.03
Total Rating**	Predictability	-0.30	0.03
Leave Behaviour***	Interaction	0.20	0.14

Table 28: Relationships between performance criteria and Insight scale

Average Handling Time

Orderliness is negatively correlated with call handling times (*r* = -0.28, *p* = 0.08). *Orderliness* measures the need for order and structure. It embodies reliability, responsibility, conscientiousness and constraint. The higher the individual's *Orderliness* score the more likely they are to have shorter handling times. All other things being equal, in particular provided there is no problem with issue resolution, low call times are desirable. We took three measures of call handling time, the average for each of the three previous months. We could have used each of the three months as the criterion, or an aggregate such as the mean, or the best of the three. The relationships below are presented for the best of the three months. However, similar relationships are found for the remaining individual months and for the mean average handling time across three months.

Call Coaching

There is strong evidence that scores on the *Physical* scale are negatively related to call coaching scores (*r* = -0.44, *p* = 0.00). The *Physical* scale assesses the importance of working outside and being involved in physical work. As is the case with the handling time criterion, we have used the best coaching score over the three months as the criterion. However, similar patterns are evident across the three individual months and the average of the three months.

Leave and Absenteeism Data

We had a number of performance indicators in this area. The *Interaction* scale was positively associated with leave behaviour (the strongest relationship was *r* = 0.20, *p* = 0.14). The *Interaction* scale assesses the importance of interacting with and helping others in the work environment. Observed correlations against leave data fall into the category of substantive but non-significant, that is, the sample size was inadequate to detect correlations of this magnitude at a significance level of 5%. However, we can note that across three of the four measures of leave behaviour, the *Interaction* scale demonstrated substantive correlations.

Insight-Contact center competency relationships

Communication is conveying to, seeking and receiving from others, information in a clear, positive and sensitive manner. The Insight scale of *Predictability* is negatively correlated with communication (r = -0.30, p = 0.02). The *Predictability* scale measures the importance of working in a stable, supportive, well-organised workplace with secure employment.

Customer Service Commitment is discovering and meeting internal and external customer's needs and offering fit-for-purpose solutions. The Insight scale of *Predictability* is negatively correlated with customer service commitment (r = -0.23, p = 0.08). The *Predictability* scale measures the importance of working in a stable, supportive, well-organised workplace with secure employment.

Detail Consciousness is attending to the detail and order/correct procedure of a task and ensuring accuracy and completion. The Insight scale of *Verbal Reasoning* is positively correlated with Team Leader Detail Consciousness ratings (r = 0.25, p = 0.06). *Verbal Reasoning* is a measure of the level of competency a person has with written language, spelling and meaning of words.

Flexibility is being prepared to modify your thinking and supporting the business by accepting changes. The Insight scale *Work Pressure* scale is correlated positively with flexibility (r = 0.29, p = 0.03), indicating that the higher the work pressure score, the higher the managerial flexibility rating. The *Work Pressure* scale assesses the importance of doing work that requires a high level of effort and commitment.

Self-Management is about knowing, anticipating and managing personal behaviour and emotions in all situations. The Insight *Work Pressure* scale is correlated positively with managerial ratings of Self Management (r = -0.20, p = 0.13). The work pressure scale assesses the importance of doing work that requires a high level of effort and commitment.

Tardiness is being late for work or back from a break. The *Verbal Reasoning*, *Numerical Reasoning* and *Logical Reasoning* scales were all positively correlated with tardiness, and as a result, so is overall ability (r = 0.29, p = 0.03). The higher your total ability scores the higher your tardiness rating.

Total rating is a composite that was the sum of all of the individual competency ratings. *Predictability* correlated negatively with this rating (r = -0.30, p = 0.03). The *Predictability* scale measures the importance of working in a stable, supportive, well-organised workplace with secure employment.

Employees who perform well in call centre environments score higher on ability scales, are more orderly, enjoy work that requires effort and commitment, have a lower preference for outdoor physical work, and a show lower need for stability, security and predictability at work.

Case Study Two

Job Performance

An analysis was conducted of the job descriptions of 65 staff from 9 job categories in a contact centre environment of a large corporation. Key performance indicators and key job responsibilities were derived from each and analysed by industrial and organisational psychologists. The results were grouped under 13 competency headings.

Five-point behaviourally anchored rating scales were then developed for each of the 13 competencies identified in Table 29 below. The employees were rated by their immediate managers on the five point scales with regard to how they performed against each of the 13 competencies. The competencies identified through the job description analysis and against which staff members were rated are presented in Table 30 below.

Corporate Competencies
Total Performance
Promoting the corporate brand
Being supportive of other team members
Maintaining an achievement orientation
Being sales focused
Providing outstanding customer service
Accuracy
Efficiency
Timeliness
Problem solving ability
Negotiation skills
Business understanding
Proactive & uses initiative
Conscientiousness

Table 29: Corporate competencies for criterion related validation

Spearman rank order correlations were computed between the manager ratings of job performance on the 13 competencies and employee profiles on Insight. The results are presented in Table 30 below.

Competency	Best Predictor	*r*	*p*
Total Performance	Need for Interaction	-0.18	0.16
Promoting the Corporate Brand	Physical	-0.34	0.01
Collegial Support	Tolerance	-0.19	0.13
Achievement Orientation	Numerical	0.25	0.04
Customer Service	Physical	-0.24	0.05
Accuracy	Self Confidence	0.22	0.07
Efficiency	Interaction	-0.20	0.11
Timeliness	Physical	-0.22	0.08
Problem Solving	Logical	0.35	0.00
Negotiation Skills	Logical	0.20	0.11
Business Understanding	Total Ability	0.34	0.01
Proactive	Physical	-0.36	0.00
Conscientiousness	Physical	-0.20	0.12

Table 30: Correlations between scale scores and corporate competencies

Behavioural Characteristics

This study investigated whether behavioural patterns identified on Insight were identifiable by managers. Managers rated these workers on each of the Insight constructs using corresponding five point Likert scales. For example, a five point rating scale was constructed for extroversion, and managers rated staff on extroversion, one of the seven scales in the My Personal Styles section of Insight. Spearman rank order correlations were calculated between the management ratings and staff scores on the respective Insight scales.

Spearman rank order correlations were computed between the manager ratings of job performance on the 13 competencies and employee profiles on Insight. The results are presented in the table below.

Scale	R	*n*	*p*
Extroversion	0.33	64	0.01
Orderliness	0.12	64	0.34
Openness	0.11	64	0.37
Teamwork	0.14	64	0.28
Tolerance	0.20	64	0.12
Competitiveness	0.20	64	0.11
Self Confidence	-0.06	64	0.62
Physical	0.22	63	0.09
Security	-0.02	64	0.89
Work Pressure	-0.01	64	0.96
Job Autonomy	0.03	64	0.82
Work Complexity	0.02	64	0.90
Interaction	0.07	64	0.58

Table 31: Correlations between scale scores and managerial behaviour rating

Normative Base

Norms are comparison groups by way of which we interpret a score. They are important because the nature of psychometric assessment means that we have no way of interpreting what a score on a test scale means without reference to a comparison group. For example, what does a score of 15 out of 30 mean with regard to ability? What does a score of 0 or 30 mean? By itself such data are of limited use. We give them meaning by saying what score represents relative to other people who have completed the test.

These comparison groups are known as *norm groups*. Clearly it is more meaningful when individuals are compared against norm groups comprised of individuals of similar ability to the individual completing the test. For example, you would not compare an engineer against a norm group of school children. It would, however, be very useful to compare the score of an engineer against other tertiary qualified individuals. It would be even more useful to compare an engineer against a norm group of engineers.

In assessing the suitability of the norm group, take into account the similarity between the individual's background and the backgrounds of those in the comparison group, and, more importantly, consider whether the background of the individuals comprising the norm group is similar to the background of people in similar roles to the role in question.

Norm Characteristics

The norm group for Selector Insight consists of 6889 individuals who completed Insight for job applications or organizational development over the period March 2002 to July 2006. It is on this large sample the ability analyses were based.

The Personal Styles and Work Preference analyses have been based on 755 individuals, 332 (44%) were male and 423 (56%) were female; 302 (40%) had a highest qualification at the tertiary level, and 453 (60%) had a highest qualification lower than at a tertiary level. This data was collected over the period March 2002 to March 2003.

Development sample

The sample used in the development of Insight was comprised of 503 participants. The factor tables presented in this manual are from the development sample. Of the 503 participants, 267 were male and 236 were female. English was the first language of 96% of the sample. The average age of the sample was 37.

References

Barrick, M. R., Mount, M. K. & Judge T, A. (2001) Personality and Performance at the Beginning Of The New Millennium: What Do We Know And Where Do We Go Next? International Journal of Selection and Assessment, 9, 9-30.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238-246.

Bollen, K. (1989), Structural Equations with Latent Variables, John Wiley & Sons, New York, NY.

Borman, W. C. & Motowidlo, S. J. (1997). Task performance and contextual performance: The meaning for personnel selection research. Human Performance, 10, 99-109.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 445-455). Newbury Park, CA: Sage

Byrne (1998) Structural Equation Modeling With Lisrel, Prelis, and Simplis: Basic Concepts, Applications, and Programming (Multivariate Applications Book Series) Lawrence Erlbaum Associates.

Chernyshenko, O. S., Stark, S., Crede, M., Wadlington, P., & Lee, W. (2003, April). Improving measurement of job attitudes: The development of the IJSI. Paper presented at the 18th annual conference for the Society of Industrial and Organizational Psychologists. Orlando, FL.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255.

Derogatis, L.R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. Psychological Medicine, 13, 595-605.

Fletcher, R. B. (2006). A Psychometric Analysis of the Ability Measure of Selector Insight. *Confidential Report for Selector Limited.*

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage

Hanisch, K. A., & Hulin, C. L. 1991. General attitudes and organizational withdrawal: An evaluation of a causal model. Journal of Vocational Behavior, 39, 110-128.

Hulin, C.L., & Judge, T.A. 2003. Job attitudes. In Borman, W.C., Ilgen, D.R., & Klimoski, R.J. (Eds.) Handbook of Psychology, Volume 12 (pp.255-276). Hoboken: John Wiley & Sons.

International Personality Item Pool (2001). A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (http://ipip.ori.org/). Internet Web Site.

Jöreskog, K. & Sörbom, D. (1986). LISREL 6: Analysis of linear structural relationships by maximum likelihood and least square methods. Mooresville, IN: Scientific Software.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User's reference guide. Chicago, IL: Scientific Software.

Kline, P. (2000) A Psychometrics Primer. Free Assn Books. Mulaik, S. A., James, L .R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430-445.

MacCallum, R. (1998). Commentary on quantitative methods in I/O psychology. The Industrial-Organizational Psychologist, Vol. 35, No. 4.

Roussos, L., & Stout. W. (1996). A multidimensionality-based DIF Paradigm. *Applied Psychological Measurement, 20*, 355-371.

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 Findings. Psychological Bulletin, Vol. 124, No. 2, 262-274.

Shealy, R., & Stout, W. (1993a). A model-based standardization approach hat separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 85,* 159-194.

Shealy, R., & Stout, W. (1993b). An item response theory model for test bias and differential item functioning. In Holland, P. W. & Wainer, H. (Eds.) *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. Multivariate Behavioural Research, 25, 173-180.

Stout. W., & Roussos. L. (1996). *SIBTEST Users Manual* (2nd ed) [Computer program manual]. Urbana-Champaign: University of Illinois, Department of Statistics.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.