Test Partnership

Insights Series

Technical Manual

2017

Test Partnership

First published March 2017

Test Partnership Ltd
http://www.testpartnership.com

Table of Contents

Tables, Figures and Equations

Section 1: Background

Overview

This section of the manual provides background and introductory information on the Insights Series and aptitude testing more generally.

What is the Insights Series?

The Insights Series is a set of three computer adaptive aptitude assessments designed for selection and assessment. They are designed to be used for selection at any job level and are applicable for most job roles. Each test has a maximum time limit of 15 minutes, and typically takes around 12 minutes to complete. Each assessment has a range of available norm groups, allowing for benchmarking at any job level from entry level staff to senior executives.

The Insights Series comprises three reasoning assessments, which can be used individually or in combination. These assessments are:

Insights Verbal
Insights Numerical
Insights Inductive

Aptitude tests, such as the Insights Series, measure specific cognitive abilities, known as aptitudes. When aggregating the scores of these aptitudes, a measure of general cognitive ability is created, improving the predictive validity of any one of the assessments used in isolation.

More details about each of the three assessments, their psychometric properties and additional features are provided throughout this technical manual.

Benefits of Aptitude Testing

Aptitude tests are among the most commonly-used assessment tools. Some of the benefits of aptitude testing are as follows:

1. *Validity and ROI:* Cognitive ability tests, such as those included in the Insights Series, are the strongest predictors of job performance known (Schmidt & Hunter, 1998). Research suggests that almost half the variance in complex job performance can be attributed to cognitive ability, making cognitive ability the single most important variable when predicting employee performance (Bertua, Anderson & Salgado, 2005). This high level of predictive validity, combined with the low cost of aptitude testing provides a significant return on investment (ROI). Moreover, aptitude tests show significant incremental validity when used in combination with other selection tools, such as personality questionnaires, situational judgement tests and interviews, increasing ROI even further.

2. *Ease of Administration:* Online administration of the Insights Series to large volumes of candidates is quick and straightforward. A single administrator can invite thousands of candidates in a single click, automating the process and reducing the administrative burden. Selecting successful candidates for the next stage is equally simple, as candidates can be rank ordered by score, identifying those which meet or exceed that standard. When compared to time and labour intensive selection tools, such as interviews, aptitude tests increase the ease of administration considerably, making them far more suitable early stage recruitment tools.

3. *Administration Time:* Compared to virtually all other employee selection tools, aptitude tests require less administration time. This is especially true of the Insights Series, which employ computer adaptive testing to further reduce test administration time. On average, Insights tests can be completed within around 12 minutes, saving time for both candidate and client. Combined with flexible online administration, aptitude tests are among the most convenient selection tools available today.

4. *Customisation and Relevance:* Many aptitude tests, such as verbal and numerical reasoning tests, can be designed with specific topics in mind, allowing for a high degree of customisation. Organisations can request bespoke assessments with question topics matching the industries they operate in, increasing the face validity of the assessment.

5. *Robust and Built on Research:* Cognitive ability and aptitudes are perhaps the most researched assessments in the field of psychology. Almost 100 years of research has consistently supported their validity, making aptitude testing one of the most supported evidence based selection tools known.

Applications of the Insights Series

The three primary applications for the Insights Series are as follows:

1. *Applicant Sifting:* The primary use for the Insights Series is early stage candidate sifting. The benefits of the Insights Series are maximised when used early in the recruitment process, to refine the candidate pool into a more manageable shortlist. As an online tool, the Insights Series can be combined with other psychometric tools, particularly personality questionnaires and situational judgement tests.

2. *Assessment Centres:* The Insights Series can also be used at assessment centres or following an interview. Ability tests can be completed under supervised conditions during an assessment centre or following an interview, thus guaranteeing the test taker's identity. Alternatively, the shorter verification test can be used to verify an initial score from the sifting process, ensuring the validity of the original score.

3. *Promotion and Succession:* The Insights Series can also provide objective data to aid in promotion and succession decisions. As job complexity increases, cognitive ability becomes more integral to performance, and thus aptitude testing becomes a stronger predictor of performance. Therefore, aptitude tests provide useful insight, contributing to promotion and succession decisions alongside other sources of information.

Advantages of the Insights Series over other aptitude tests

The Insights Series uses some of the most powerful advancements in psychometrics, providing numerous advantages over most traditional aptitude tests:

1. *Increased Reliability:* The Insights Series boasts higher levels of reliability than most currently-available aptitude tests of similar length. This is because the Insights Series employs computer adaptive testing (CAT), maximising the reliability of the assessments far beyond what is possible with fixed form, randomised or linear-on-the-fly (LOFT) testing.

2. *Faster Testing:* As reliability is maximised through CAT, the Insights Series can be shorter than most traditional assessments. This results in time saving for candidates and clients, which is especially important during assessment centres and supervised testing scenarios.

3. *Greater Security:* Unlike tests with a fixed set of questions, the Insights Series uses large and frequently updated item-banks. This ensures that candidates are provided with individual test experiences, making it significantly more difficult for candidates to cheat. Even if dishonest candidates attempt to share questions, the odds of those same questions appearing in another candidate's assessment are slim, removing the advantage.

4. *Flexible:* Item banks provide other benefits to the Insights Series in that questions can be easily added or removed. This ensures that item banks can continually grow and evolve, further increasing the reliability and security of the assessments. Similarly, items can be removed from the item banks without disruption, keeping the questions current and up-to-date.

5. *Candidate Experience:* Adaptive tests target the difficulty of the assessment to the ability of the candidate. This ensures that candidates do not receive overly easy or difficult questions, reducing the chance of candidates losing interest or becoming overwhelmed by the difficulty of the questions. Instead, everyone is given an engaging test experience, regardless of their individual level of ability.

6. *Fairer:* The Insights Series employs dynamic item bias detection software, investigating item banks for evidence of item bias on a continual basis. If a question is found to disadvantage a demographic, this question is automatically flagged for review, allowing administrators to review and if necessary remove the question. This ensures that bias is constantly reviewed, and item banks continue to produce fair assessments for all candidates.

Overview of the Insights Series

The Insights Series is a collection of online aptitude tests which can be used to assess verbal, numerical, and abstract reasoning ability. Or if used in conjunction with one another, they can provide a measure of general cognitive ability.

Insights tests can be used as an early screening procedure or adopted at later stages in the recruitment process. Insights Tests can be administered in either supervised (proctored) or unsupervised settings.

Each of the Insights tests has an optional verification test. This is a short form of the original assessment that can be administered to candidates who were previously assessed unsupervised, as a confirmation of their test scores. The verification feature allows clients to assess large volumes of candidates in a remote unsupervised capacity, with confidence in the authenticity of the initial test scores.

Use of the Insights Series can be managed easily using our online assessment platform. The Insights Series consists of three tools:

Insights Numerical – 15 Questions (Maximum time limit: 15 Minutes)

Insights Verbal - 20 Questions (Maximum time limit: 15 Minutes)

Insights Inductive - 15 Questions (Maximum time limit: 15 Minutes)

The corresponding verification tests for each Insights Test is approximately 60% of the length of the original assessment:

Verification - Insights Numerical – 9 Questions (Maximum time limit: 9 Minutes)

Verification - Insights Verbal - 12 Questions (Maximum time limit: 9 Minutes)

Verification - Insights Inductive - 9 Questions (Maximum time limit: 9 Minutes)

Unlike traditional tests with a fixed set of questions, Insights tests are adaptive. Adaptive tests increase in difficulty when questions are answered correctly and decrease in difficulty when answered incorrectly, tailoring the test's difficulty to the performance of the candidate.

Research suggests that questions equivalent to a candidate's ability provide the most effective measures, making adaptive tests considerably more reliable than traditional tests of comparable length. Each aptitude test draws from an item bank containing hundreds of pre-calibrated questions, allowing the testing algorithm to select the questions most targeted to a candidate's ability.

Constructs Measured by the Insights Series

The Insights Series

Numerical, verbal, and inductive reasoning are considered facets of cognitive ability. There is a vast amount of evidence regarding the relationship between cognitive ability and job performance. Research suggests that cognitive ability is the single best predictor of job performance, especially within complex roles.

Due to the predictive power of cognitive ability, it can be safely assumed that any true measure of cognitive ability can predict job performance. Included within general cognitive ability are numerical, verbal, and abstract reasoning facets. When the full Insights Series is used to assess a candidate, this provides an indication of their general cognitive ability.

The Insights Numerical Reasoning Test

The Insights Numerical Reasoning Test provides a measure of how well a candidate can analyse and interpret numerical information and perform calculations based on this information.

Candidates will be required to interpret and analyse numerical information that is presented in the form of either a graph, a table, or a short passage. To correctly work out the answer, candidates will be required to calculate basic numerical equations that are approximately equivalent to GCSE level.

The questions are constructed in such a way that candidates will first have to consider how the problem can be solved and to identify the correct information necessary to do so, before performing the calculations. Candidates will be allowed the use of a calculator, as this reduces the need for supervision. The Insights Numerical Reasoning Test contains 15 questions and has a maximum time limit of 15 Minutes.

The Insights Verbal Reasoning Test

The Insights Verbal Reasoning Test provides a measure of how well a candidate can evaluate, reason, and conceptualise with words and sentences.

Candidates will be required to evaluate, reason, and conceptualise using written information presented in a short passage. To correctly work out the answer, candidates will be required to evaluate the validity of conclusions deduced from the passage of written information, based on the following options:

Definitely true – The statement is definitely true beyond a reasonable doubt, based solely on the information in the passage.

Probably true – The statement is more likely to be true than false, but not definitely true beyond a reasonable doubt, based solely on the information in the passage.

Insufficient information - There is not enough evidence to make a decision based solely on the information provided in the passage.

Probably false - The statement is more likely to be false than true, but not definitely false beyond a reasonable doubt, based solely on the information in the passage.

Definitely false - The statement is definitely false beyond a reasonable doubt, based solely on the information in the passage.

The Insights Verbal Reasoning Test contains 20 questions and has a maximum time limit of 15 Minutes.

The Insights Inductive Reasoning Test

The Insights Inductive Reasoning Test provides a measure of how well a candidate can think logically, identify patterns, and apply abstract problem solving.
Candidates will be required to apply abstract problem solving to identify an underlying pattern in a logical sequence of diagrams. To identify the correct answer, candidates will need to identify the missing diagram in the sequence from a range of possible options.

Each question has five possible answers, only one of which is the correct answer based on the pattern which underlies the sequence. The Insights Inductive Reasoning Test contains 15 questions and has a maximum time limit of 15 Minutes.

Section 2: Theoretical Framework

Overview

This section of the manual provides a broad overview of the theoretical and empirical frameworks behind the Insights Series. This section will provide an overview the following areas:

Cognitive ability and aptitude testing

Item Response Theory (IRT)

Computer Adaptive Testing (CAT)

Cognitive Ability and Aptitude Testing

When individual scores from the Insights Series are aggregated, this aggregated score provides a measure of general cognitive ability. When completed individually, specific cognitive abilities, known as aptitudes are measured. The literature regarding general cognitive ability, aptitudes and the differences between the two, are discussed below.

General Cognitive Ability (GCA)

General cognitive ability (GCA) can be best described as “a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience.” (Gottfredson, 1997). GCA was discovered by Charles Spearman in the early 1900s after identifying positive relationships among scores on different cognitive tasks. For example, high performers on numerical reasoning tests also tended to perform well on measures of seemingly unrelated abilities, such as verbal or logical reasoning. Spearman hypothesised that an underlying factor accounted for this phenomenon, a general mental ability which influences performance across every cognitive domain. After inventing the statistical technique of factor analysis, spearman discovered his hypothesised underlying factor, the general cognitive ability factor, or *g* for short.

Since its discovery, the validity of general cognitive ability has shown itself to be ubiquitous, predicting performance on a wide range of outcomes. Significant associations with a huge range of variables have been identified including (but not limited to): educational achievement (Deary, Strand, Smith & Fernandes, 2007); job and training performance (Schmidt and Hunter, 1998); physical fitness (Etnier, Salazar, Landers, Petruzzello, Han & Nowell, 1997); and occupational attainment (Schmidt & Hunter, 2004). The strong predictive validity of general cognitive ability tests explains their widespread use in employee selection and assessment. Many academics agree that cognitive ability tests should be considered the primary selection tool, with other selection tools considered supplementary (Schmidt & Hunter, 1998).

Although ample research supports the existence, validity, and utility of GCA, psychologists are divided over the finer details. For example, the origins, heritability, and malleability of GCA are still areas of contention, leaving many important questions unanswered.

Additionally, competing models of cognitive ability have been proposed which do not include GCA, such as Gardner's theory of multiple intelligences (Gardner, 1983) and Sternberg's triarchic theory of intelligence (Sternberg, 1985). These alternative models of cognitive ability however, have received significant criticism from the academic community, and have largely been discredited in favour of GCA based models (Gottfredson, 2003; Waterhouse, 2006).

Aptitude Testing

The Cattell–Horn–Carroll theory of cognitive abilities holds that specific aptitudes, broader abilities and GCA stem from a linear hierarchy (Schneider & McGrew, 2012). Aptitudes are considered partial measures of GCA, loading onto GCA to varying degrees. Additionally, lower order factors may also exist in between GCA and specific aptitudes, such as fluid intelligence. The theory holds that GCA is a superordinate psychological construct, and thus can be measured only indirectly through the aggregations of lower order factors. This model highlights the interrelatedness of aptitudes to one another, and to GCA, providing a comprehensive explanation of specific and general human cognitive abilities. Currently, the Cattell–Horn–Carroll theory of cognitive abilities is the leading academic model of cognitive ability, accepted by a large proportion of experts in the field.

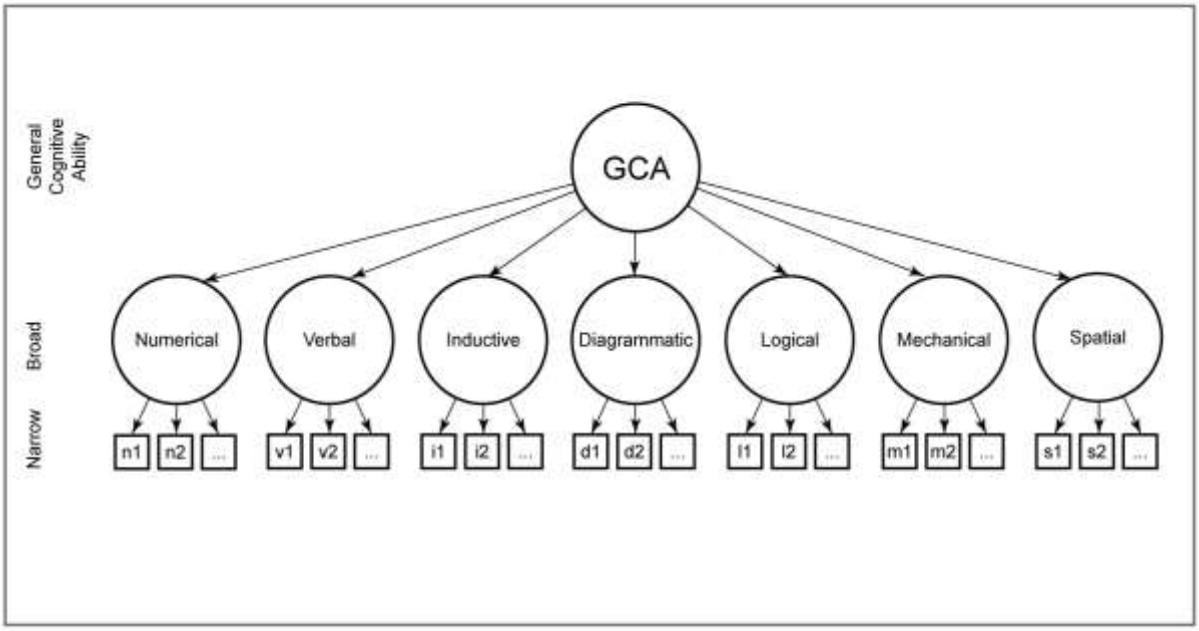

Figure 1: GCA and Aptitude Hierarchy

As the name implies, GCA is an extremely broad and general psychological construct, and cannot be measured optimally with a single aptitude test. Instead, a range of aptitude tests should be employed and aggregated, creating a holistic picture of a candidate's ability. As a superordinate construct, GCA is partially measured by every aptitude, and its validity is increased through the aggregation of more aptitude test scores. This aggregation filters out the aptitude specific variance, resulting in a pure measure of GCA. The more aptitude tests employed, the more accurately GCA is measured, and thus the more powerful the predictive power of the scores in an employee selection setting.

> *"The purely empirical research evidence in I/O psychology showing a strong link between GCA and job performance is so massive that there is no basis for questioning the validity of GCA as a predictor of job performance".*
>
> - Schmidt, F. L. (2002)

GCA is the most predictive psychological construct known when predicting employee performance (Hunter & Hunter, 1984; Schmidt & Hunter, 1998). Aptitude and GCA tests outperform all other selection tools when predicting employee performance, especially in complex professional / managerial / technical roles (Bertua, Anderson & Salgado, 2005). Research shows that almost 50% of variance in complex job performance is attributable to GCA (Bertua, Anderson & Salgado, 2005). Although GCA's predictive validity is highest for complex work, aptitude tests and GCA measures are still among the most useful predictors of performance in moderate and low complexity work, matching or outperforming other commonly used selection tools (Hunter & Hunter, 1984).

GCA is also a strong predictor of training performance, an important variable informing employee selection decisions (Bertua, Anderson & Salgado, 2005; Schmidt & Hunter, 1998). The greater the training performance of employees, the greater return on investment generated by training programmes and initiatives. Candidates with higher scores on aptitude tests and GCA measures are more likely to benefit from training programmes than candidates with low scores. GCA is strongly associated with the ability to learn, retain and apply new information, resulting in higher performance on training programmes and thus a greater return on investment for employing organisations.

Most important, is the incremental validity provided by aptitude and GCA measures when combined with other selection tools (Schmidt & Hunter, 1998). Combining these assessments with situational judgement tests (SJTs), interviews, and personality questionnaires yields additional predictive power, further improving selection process validity. This incremental validity is maximised when aptitude and GCA tests scores are combined with tools measuring unrelated psychological constructs, such as personality traits (Schmidt & Hunter, 1998). Using these assessments in combination rather than isolation provides a broader picture of candidate's potential, improving the quality of selection decisions.

The use of aptitude tests in employee selection may also reduce staff turnover, improve employee retention, and increase average employee tenure, (Mount, Witt & Barrick, 2000). Candidates scoring low on measures of GCA are at risk of feeling overwhelmed by a roles cognitive demands, especially in complex professional / managerial / technical work, increasing employee turnover. Other low-scoring candidates may meet the cognitive demands of their current role, but could not cope with a more senior position, leading to short tenures. High performers however, meet the roles cognitive demands regardless of seniority, resulting in longer tenures and lower dropout rates.

Due to the substantial utility of aptitude tests and their low cost, their use in employee selection can result in significant return on investment. Aptitude tests are typically low-cost and require little administrative effort, especially when completed online. In contrast, employment interviews may last over 45 minutes and must be conducted by one or more trained and experienced members of staff. Employee performance gains are directly proportionate to increases in selection process validity, significantly improving following the addition of aptitude and GCA measures.

Summary

General cognitive ability, as measured by a range of aptitude tests, is the most powerful single predictor of job performance known. GCA is most predictive of performance in complex work, but is still a powerful predictor of performance in moderate and low complexity work. Combining GCA measures with additional selection tools further increases the predictive validity of any selection process. GCA measures complement SJTs, personality questionnaires, and interviews, significantly improving the quality of selection decisions.

Item Response Theory (IRT) and the Rasch Model

The Insights Series employs the Rasch model, a statistical model which parameterises the difficulty of administered questions when estimating a candidate's ability. The Rasch model and other item response theory methods are modern approaches to test development, affording greater flexibility than classical methods. The differences between item response theory, the Rasch model and classical test theory are discussed below.

Item Response Theory vs. Classical Test Theory

tTraditional assessments either implicitly or explicitly rely on classical test theory (CTT) methods during construction and scoring. CTT can be summarised using the following equation:

$$\mathbf{X} = \boldsymbol{T} + \mathbf{e}$$

X= Individuals observed score on the test
T = Individuals true score on the test
e = Measurement error

Equation 1: Classical Test Theory

Although simple, CTT places significant constraints on assessments. CTT assumes that true score and error are the only sources of variation in test score. To ensure this, all candidates must receive a fixed set of questions, avoiding the influences of additional variables on observed scores. This creates the possibility of cheating, especially when administering questions online. Fixed test forms also reduce the flexibility of assessments, limiting assessments to specific difficulty levels. Because the difficulty levels of fixed form

assessments are static, they cannot measure candidates at different ability levels with equal precision.

Item response theory (IRT) methods however, seek to improve measurement by employing complex statistical procedures to scoring. Most importantly, IRT employs an item difficulty parameter when estimating a person's ability, weighting scores based on the difficulty of the questions administered. This frees assessments from requiring fixed forms, allowing candidates to receive a unique set of pre-calibrated questions from an item bank. Large item banks provide a robust solution to cheating, ensuring that different candidates receive unique testing experiences, reducing their ability to share answers and compromise test security.

Item Response Theory Models

IRT encompasses a range of statistical models for estimation. Most commonly, one of four models are used, each with different parameters and / or underlying statistical assumptions. These include:

The One-Parameter Logistic model (1PL): The 1PL model parameterises question difficulty, weighting scores based on the difficulty of the questions administered. When calculating scores, the number of correct answers and the difficulty of the questions both contribute, equating candidate's scores regardless of the specific questions provided. If the data do not fit the 1PL model, IRT practice suggests that the 2PL model should be tested.

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}$$

Equation 2: The 1PL model

The Two-Parameter Logistic model (2PL): The 2PL model also employs a difficulty parameter, but also an item discrimination parameter, this refers to a questions ability to discriminate between high and low performers. This results in higher discriminating questions receiving a greater weighting when calculating scores. The additional parameter increases the sample size requirements for item calibration compared to 1PL. If the data do not fit the 2PL model, IRT practice suggests that the 3PL model should be tested.

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Equation 3: The 2PL model

The Three-Parameter Logistic model (3PL): Alongside the difficulty and item discrimination parameters, the 3PL model employs a guessing parameter. If certain questions are easily guessed, candidate scores can be adjusted based on the "guessability" of each administered question. Although more complex models have been created, the 3PL model is generally considered the most complex commonly

used model. Due to its complexity and the requirement to calibrate three separate parameters, it requires larger sample sizes than 2PL.

$$P_i(\theta) = C_i + (1 - C_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}$$

Equation 4: The 3PL model

The Rasch Model: The Rasch model is a special case of the 1PL model, with different mathematical and practical considerations. When using traditional IRT modelling, if low discriminating or guessable questions are identified, the model must be changed to account for these variations. With the Rasch model, the model is considered superior to the data, and so low discriminating and guessable questions are removed from the assessment, while retaining the model.

Advantages of the Rasch model

To score the Insights Series, the Rasch model was selected over traditional IRT scoring procedures. There are numerous advantages afforded by the Rasch model, including the following:

1. *Required sample size*: 3PL modelling may require a minimum of 1,000 participants per item to ensure proper calibration (Tang, Way & Carey, 1993). The Rasch Model however, requires a minimum of 30 participants for stable calibration, with over 250 preferred for high-stakes testing (Linacre, 1994). Smaller sample size requirements result in faster item bank development, and ease of item bank growth.

2. *Added flexibility*: A unique advantage of the Rasch model, is that raw scores are a sufficient statistic (Wright, 1989b). This means that Rasch calibrated tests can be administered conveniently both online and via paper and pencil format. Other IRT models however, require advanced scoring software for paper and pencil testing, or will not be applicable at all.

3. *No assumed guessing:* If a low performing candidate correctly answers a difficult question, the 3PL model often considers this guessing, awarding the candidate less credit. In practice, that candidate may or may not have guessed, with no objective method of identifying true guessing. Similarly, if the candidate did guess, they may have employed informed guessing, deserving some credit. The Rasch model gives candidates the benefit of the doubt, rather than penalising low performers for correctly answering harder questions.

4. *Specific objectivity:* The Rasch model holds to a standard of objective measurement akin to measurement in the physical sciences (Bond & Fox, 2015). In the Rasch model, ability estimates of people are invariant over the specific items used, and item difficulty estimates are invariant over the specific people used to calibrate them. This is known as specific objectivity, and of the IRT models only the Rasch model guarantees this property.

5. *Equally effective:* Research suggests that the additional parameters required by the 2 and 3 parameter logistic models do not improve the accuracy of the assessment compared to the Rasch model (Anderson, 1998; DeMars, 2001; Pelton, 2002). This suggests that the disadvantages of using complex models are not offset by increases in accuracy, supporting the use of the simpler Rasch model.

Summary

Designing assessments using IRT affords numerous advantages over CTT, especially regarding cheating and test security. Of the available IRT models, the Rasch model was chosen to design, calibrate, and score the Insights Series. The Rasch model holds practical advantages over other IRT models, without any significant drawbacks.

Computer Adaptive Testing (CAT)

The Insights Series employs computer adaptive testing to improve reliability, item security and candidate experience. Computer adaptive tests administer the optimal set of items to each candidate based on their unique level of ability. The typical CAT procedure, along with the relative advantages and disadvantages of CATs are discussed below.

Computer Adaptive Testing Procedure

IRT offers an additional benefit over traditional CTT assessments by permitting creation of computer adaptive tests (CAT). Unlike fixed form assessments, CATs adjust their difficulty to match the candidate's performance, increasing in difficulty with correct answers, and reducing in difficulty with incorrect answers. After each question is administered, the ideal next question for each candidate is selected from the item bank, ensuring that candidates are issued with the most informative questions available based on their level of ability. This allows the CAT to hone in on the candidate's level of ability, increasing the accuracy of the assessment considerably compared to a fixed form test. Research shows that CATs can reduce test length by 50% or more, while maintaining an equal or greater level of reliability compared to equivalent fixed form tests (Linacre, 2000).

CATs follow a multistage process, starting with a pre-calibrated item bank and culminating in satisfaction of a stopping rule. The typical CAT process stages can be seen below:

1. *Calibrated Item Bank:* In the first stage, a pre-calibrated item bank is accessed which comprises items with known psychometric properties. The calibration of item banks (the process of discovering the item's psychometric properties) requires significant investment in both time and resources. Once calibrated, item banks should contain large numbers of thoroughly-researched items, ideally with a wide spread of item difficulties. This ensures that candidates of all abilities have questions targeted to their level, readily available within the item bank.

2. *Starting Rule:* Without more information regarding the candidate's level of ability, a starting item of average difficulty level is usually selected from the bank. Certain CATs may utilise available information, such as previous test results, to estimate the

candidate's likely level of ability, administering a first item close to that estimate. Due to the adaptive nature of CATs, the targeting of the first item need not be perfect, as the assessment eventually hones in on their level of ability as they progress.

3. *Item Selection:* As the candidate progresses through the assessment, information is gathered regarding the candidate's level of ability. This information is used to identify the most effective item for subsequent administration, which varies depending on the chosen item selection procedure. Although many selection procedures may be implemented, candidates are generally provided with more difficult questions following correct answers, and easier questions following incorrect answers.

4. *Scoring procedure:* IRT based scoring is applied based on the available information. This score will update after the administration of each question, with estimated ability scores increasing in precision following each item administration. At a minimum, this scoring procedure incorporates the proportion of correct answers and their respective item difficulties. More complex IRT models may also model guessing and item discrimination. Once a score has been generated, this score is used to inform item selection, or if the stopping rule has been satisfied, the score is retained as the candidate's final score.

5. *Stopping rule:* Several different stopping rules can be applied, such as a minimum number of questions administered or a minimum level of precision reached. Once the stopping rule has been satisfied, the test is concluded and a final score is generated. If the stopping rule has not been satisfied, the CAT returns to stage 3, selecting and administering an additional item. This iterative process continues until the stopping rule has been satisfied, ensuring the administration of enough items.

Information Provided by Computer Adaptive Testing

Research shows that the most informative questions (those contributing most to reliability) are those which candidates hold a 50% chance of answering correctly (Linacre, 2000; Van der Linden & Glas, 2000). As candidate abilities vary considerably, the optimal set of questions is candidate-specific. Fixed form assessments however, have a set difficulty which is off target for almost every candidate. Tests of moderate difficulty provide ineffective measures of high and low performing candidates, reducing reliability. However, when test difficulty dynamically matches candidate's abilities, ideal test difficulty is achieved for each candidate, increasing the assessment's reliability.

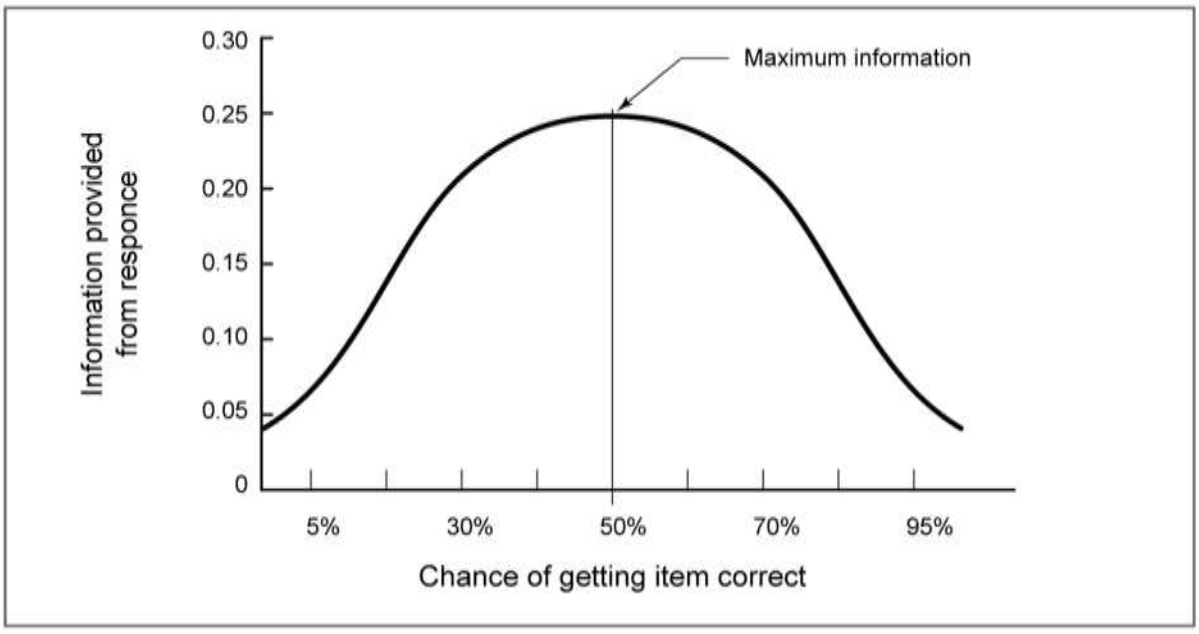

Figure 2: Information Provided Dependent on Difficulty Targeting

As shown above, with perfect item targeting, a candidate has a 50% chance of correctly answering the item, providing the maximum amount of information. Information is directly proportionate to reliability, with more informative tests offering higher levels of reliability. Because individual abilities differ, the most informative item is contextual, with harder items providing more information for high performers and easy items providing more information for low performers. Typically, fixed form assessments contain a range of item difficulties, with a small percentage of items optimally targeted to any specific candidate. CAT however, ensures that all questions are targeted to each candidate's ability, reducing the number of items required to achieve a reliable score.

Advantages and Disadvantages of Computer Adaptive Tests

Although CATs represent the most powerful advancement in psychometric testing, they are not without drawbacks. The advantages and disadvantages are listed and contrasted below:

Advantages of CAT

Increased Reliability: By ensuring that candidates receive only the most informative items in the bank, test reliability of CATs is enhanced considerably compared to other administration processes. Increases in reliability directly translate to increases in test validity, ensuring a higher level of ROI in any employee selection process.

Reduced Testing Time: Because candidates are not presented with overly easy or overly difficult items, which do little to improve reliability, administration time can be reduced without sacrificing reliability. For an equivalent level of reliability, CATs can be around half the length of their fixed form counterparts, greatly reducing administration time to the benefit of clients and candidates alike.

Increased Test Security: CATs require large item banks, ensuring that candidates receive individual testing experiences. Should dishonest candidates copy and share questions online, other candidates will be unlikely to see the same questions during their assessments. This limits the benefits of cheating, increasing the security of the assessment, and the validity of the scores.

Suitable for Different Levels: Large item banks for CATs contain questions of varying difficulty, ranging from very easy to very difficult. This ensures that a single, large item bank is sufficient for testing at any job level, from entry level staff to senior managers. This means that multiple test forms for different job levels are no longer required, ensuring that everyone is tested to the same standard.

Updating and Improving: Using item banks adds a level of flexibility unavailable to fixed form assessments. Unlike with fixed form assessments, new questions can be added to the bank and old questions can be removed without any disruption to the candidates or the tests. This ensures that question content can remain up-to-date and relevant, while also increasing the size of the item banks to improve reliability and test security.

Disadvantages of CAT

Research Intensive: Developing a large item bank for CAT requires significant resources and access to psychometric tests. Thousands, if not tens of thousands of participants are required during the trialling process to calibrate a single CAT item bank. Few organisations have the resources and capability to develop assessments on this scale.

Requires Significant Expertise: Knowledge and experience with IRT methods is rare in comparison to classical test theory. Developing a CAT item requires specially trained psychometric experts, with knowledge of psychometric methods, statistics and computer adaptive testing protocols.

Must be Online or Computer-Based: As adaptive tests dynamically select items based on candidate performance, this requires access to a computer, and almost always a stable internet connection. This makes adaptive testing unsuitable for paper and pencil testing, which is suitable only for fixed form assessments.

Anxiety Provoking: Traditional CATs administer items with a 50% chance of success, to maximise reliability. In practice, this can create a stressful testing experience for candidates, especially for high performers who expect to get most questions right. To reduce test anxiety, the Insights Series reduces the difficulty targeting from 50% success to around 70%, improving the candidate experience and reducing test anxiety (see Section 3 for more information).

Not Possible to Review Previous Questions: Generally, CATs do not allow candidates to review previous answers or go back to previous questions. Item selection is heavily dependent on responses to previous questions, and so changing previous answers will interfere with the item selection process. Instead, candidates are prevented from

going back to previous questions, keeping consistent progression and item selection throughout the assessment.

Summary

Computer adaptive testing (CAT) is the most advanced development in psychometric testing practice. CATs target item difficulty to candidates' abilities, ensuring an optimal testing experience for every candidate. CATs can employ candidate friendly selection algorithms, as well as provide greater reliability and security than any other form of psychometric assessment, especially when compared to traditional fixed form tests. Although CATs are not without drawbacks, the advantages outweigh the disadvantages considerably, prescribing their use for the Insights Series.

Section 3: Insights Features

Overview

This section of the manual details the unique features of the Insights Series and how they work. These features include:

Insights Verification

Insights Targeting

Insights Bias detection

Ongoing item bank improvement

Customisation and bespoking

Insights Verification

Insights Verification allows assessors to confirm a candidate's unsupervised score by administering a shorter, supervised assessment, which compares scores from the first assessment with the second. Should the candidate's supervised score be significantly lower than the original score, their original score is deemed "Not Verified", warranting investigation by the assessor. If the candidate's second score is not significantly lower than their original score, their original score is deemed "Verified", requiring no further investigation.

Risk of Cheating

Unsupervised testing always carries a risk of cheating, potentially threatening test security. Item banking alleviates some of this risk by negating the effects of question sharing, but even item banking cannot guarantee test security. A dishonest candidate may attempt to cheat by recruiting others to complete the tests on their behalf, creating scores unrepresentative of the candidate's true ability. Although this does not guarantee a high score, if the candidate recruits a sufficiently high performer, the candidate will progress to the next stage of the recruitment process unjustly.

How Insights Verification Works

To combat cheating attempts, the Insights Series offers an optional shorter, proctored verification assessment designed to confirm the candidates original score. Verification tests are approximately 60% of the original test's length, and returns a pass / fail result depending on whether the candidate's verification score is significantly lower than their supposed original score. The verification assessment is also adaptive, ensuring sufficient reliability and precision. Figure 3 demonstrates how verification testing works in practice.

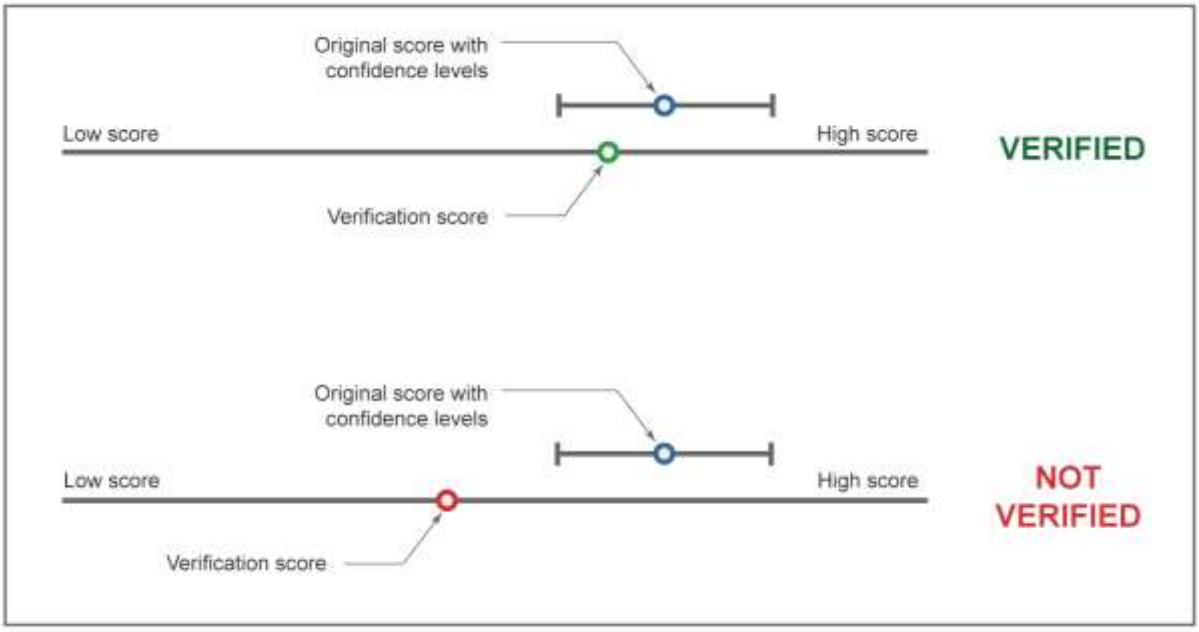

Figure 3: Insights Verification Confidence Intervals

As shown in the diagram above, verification testing employs confidence intervals when comparing the two scores. The probability of achieving a score outside the lower bound confidence interval without initial cheating is low, as most candidates will achieve a score comparable to their original assessment. However, if the candidate had indeed employed dishonest actions and was a low performer, the probability of failing the verification assessment will be considerably greater, making failed verification tests worthy of investigation.

Additionally, the expectation of verification testing may dissuade candidates from attempting to cheat in the first place. If candidates are made aware that verification testing may occur later in the selection process, it may discourage candidates from attempting to cheat, protecting the validity of the assessment. Candidates should therefore be informed prior to their initial assessment that verification testing may occur later in the selection process.

Failing a Verification Test

Should a verification test yield a failed result, this does not automatically imply the candidate is cheating. Although the candidate may have cheated, additional factors may have influenced their score, warranting caution. Many alternative explanations may account for a failed verification test, such as:

1. Physical illness or discomfort
2. Test anxiety or nervousness
3. Environmental distractions or noise
4. Assessment fatigue during assessment centres
5. Technical problems
6. Misreading instructions
7. Not having supporting materials such as a calculator

After failing verification, candidates should be asked for any alternative reasons which may explain the failed verification process. If the administrator is satisfied with the candidate's explanation it is recommended to administer an additional full-length proctored assessment. This provides candidates with another opportunity to perform, without passing judgement or accusing the candidate outright of cheating. Should the candidate fail to achieve the original required pass mark, they should not progress to the next stage of the recruitment process.

Summary

Insights verification confirms the validity of initial unsupervised scores under supervised conditions. A shorter supervised verification test is completed and compared to the original, unsupervised score. If the verification score is significantly lower than the unsupervised score, verification is failed. Candidates failing verification should re-take the full assessment, and if they no longer meet the original pass mark, should not progress to the next stage of the recruitment process.

Insights Targeting

The difficulty of the items administered during Insights tests is controlled by a specially-designed item selection algorithm known as Insights targeting. The aim of Insights targeting is to provide an optimal candidate experience, while retaining the favourable psychometric advantages of CAT.

Traditional CAT Item Targeting

Maximum information item selection - the most commonly used CAT item selection procedure - requires that candidates are administered items with a 50% chance of success under the Rasch model. Item information is directly proportional to reliability and precision, with higher item information closely related to higher overall test reliability. Psychometrically, this approach ensures the highest level of precision, accuracy, and reliability. This is because question difficulties and person abilities are perfectly matched, providing the maximum possible amount of information that the model allows. Although psychometrically optimal, in practice this approach can be problematic from a candidate experience perspective.

Taking an assessment with an almost guaranteed success rate of only 50% typically provides a suboptimal testing experience, especially for high performing candidates. Perceived performance is typically associated with the number of correct answers, not with test difficulty, meaning that candidates are unlikely to perceive their performance as high, regardless of their actual performance. Similarly, as the difficulty of CATs progressively increase as questions are answered correctly, candidates may feel penalised for answering questions correctly, causing frustration. At best, traditional CAT item selection may provide a somewhat stressful testing experience, at worst it may cause potentially high performing candidates give up mid-test or become disaffected with the selection process.

Insights Item Targeting

The Insights Series applies a candidate friendly item selection procedure, when compared to traditional CATs. Instead of administering items with a 50% chance of success, the Insights Series administers questions with a 70% chance of success. Overall, this ensures that candidates will typically answer 70% of their questions correctly, providing a more agreeable testing experience. Questions of this difficulty will appear challenging and engaging, but not overwhelming, keeping candidates interested without discouragement. Similarly, questions at this difficulty are not too easy, ensuring that candidates are not bored by the level of difficulty.

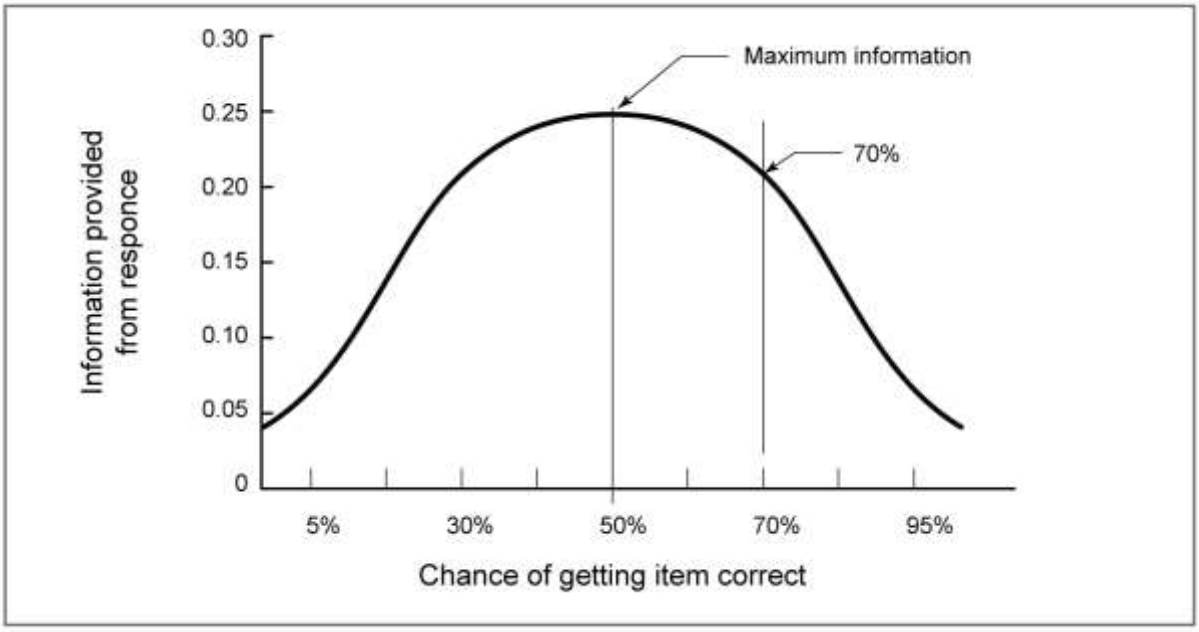

Figure 4: Information Provided by Item Targeting at 50% and 70% Chance of Success

Although this approach results in a minor reduction in precision compared to maximum information based item selection, the reduction is negligible, and the improvement in candidate experience is significant. Candidates are not discouraged or put-off by excessively difficult questions, reducing their levels of stress and anxiety. Candidates are less likely to feel they are being overly pressured by the assessment, which would normally make candidates feel uncomfortable and uneasy. Candidate experience is a vital aspect of any selection process or assessment, and Insights targeting aims to improve the candidate experience when undertaking the Insights Series.

Benefits of Insights Targeting

Research investigating the effect of CAT difficulty reduction on candidate reactions, report the following benefits (Bergstrom, Lunz & Gershon, 1992; Eggen & Verschoor, 2006; Tonidandel, Quiñones & Adams, 2002):

1. *More Motivation:* Reducing the difficulty of adaptive tests has been shown to increase the motivation of test takers. Perceived performance on CATs is unrelated to actual performance, instead only the number of correct answers appears to influence perceived performance, and thus motivation to perform. Reducing the CAT's difficulty

improves motivation to perform, making candidates more likely to try their best and less likely to drop out.

2. *Greater Satisfaction:* Higher perceived performance on CATs is directly related to improved satisfaction with the selection process. Candidates with greater selection process satisfaction are more likely to accept a job offer, recommend the job to others, report high levels of perceived organisational attractiveness, report positive attitudes towards the company and display higher work self-efficacy.

3. *Higher Perceived Fairness:* In traditional CATs, the perceived performance of high performing candidates is likely to disaccord with their actual performance, reducing perceived fairness. Reducing difficulty however, will ensure that perceived performance in high performing candidates, better matches their actual performance and thus increases perceived fairness. Candidates who view a selection process as fair are more likely to be highly committed to the organisation, show higher trust in management, and display lower staff turnover intentions.

4. *Little Loss in Precision:* Reducing the difficulty of CATs from a 50% chance of success to 70% results in only a minor reduction in precision and accuracy. This reduction in precision requires the addition of only one or two items to offset it. Yet a reduced administration time and a high level of reliability are maintained.

Although seemingly a minor detail, the reduction in overall test difficulty may result in considerable improvement to the candidate experience, benefiting both candidates and employers alike.

Summary

Ensuring that high performers are not discouraged from progressing is of paramount importance to any selection process. Insights targeting helps prevent high performing candidates from misreading their performance by reducing the average item difficulty compared to traditional CATs. This ensures that candidates are given a more agreeable testing experience, increasing motivation, satisfaction and perceived fairness, while still maintaining high level of precision and short administration time.

Insights Bias Detection

Insights bias detection automatically conducts rigorous bias detection analysis on every question within the Insights Series, across a wide range of demographics. These calculations are performed regularly, informing administrators of potential item bias as soon as it is detected. Flagged questions are then investigated and removed, preventing biased questions from unfairly disadvantaging any group.

Group Differences and DIF

Tests of group differences, as measured standardised effect sizes such as Cohen's *d,* are designed to quantify the magnitude of group differences in score. Although simple and easily calculated, standardised effect sizes are easily misinterpreted, and only limited conclusions

can be inferred from their findings. Differential item functioning (DIF) analysis is an IRT approach to detecting group bias at the individual item level, providing a more complex insight into group differences. DIF analysis evaluates the behaviour of an item across different groups, identifying whether members of specific groups have a lower probability of correctly answering the question, after controlling for overall ability. If evidence of significant DIF is found, it suggests that an item may be biased against a specific group, and that item must be flagged for review.

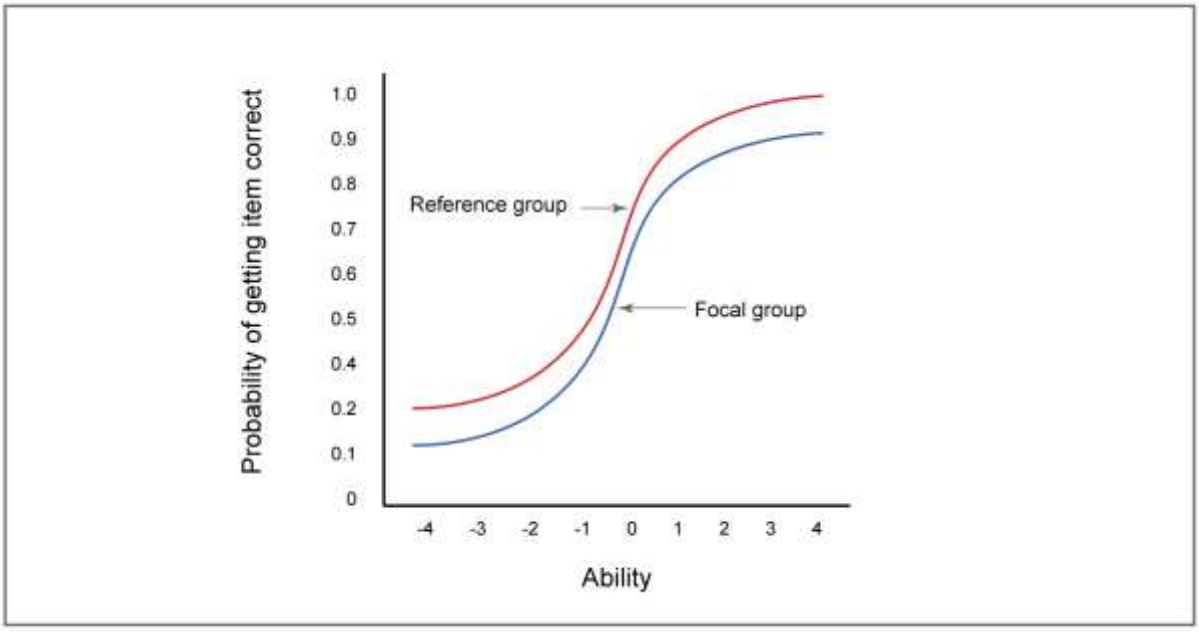

Figure 5: ICC of an Item that Demonstrates Evidence of DIF

DIF differs significantly from traditional tests of group differences in many ways. Some of the differences between traditional tests of group differences and DIF can be seen in the table below:

Group Differences (Cohen's *d*)	Differential Item Functioning (DIF)
Provides a static statistic per analysis for the entire assessment.	Provides a statistic for each individual item per analysis in the item bank.
Can be used with both classical test theory and item response theory methods	Can only be applied with item response theory methods
Cannot distinguish between item bias and genuine differences between groups.	Controls for ability differences between groups and only reports bias.
If no overall differences between groups are found, individual biased questions may be hidden.	Detects individual biased questions even if no overall differences in test score between groups are found.

Assessments that result in excessive adverse impact will need to be discontinued.	Individual items displaying excessive DIF can be removed without needing to discontinue the assessment.
Requires small / moderate sample sizes to ensure adequate statistical power.	Requires large sample sizes to ensure adequate statistical power.

Table 1: Comparisons between Tests of Group Differences and DIF

How Insights Bias Detection Works

Rather than merely running a DIF analysis only during the development stage, the Insights Series automatically runs DIF analyses on its three item banks on an ongoing live basis. This ensures that the Insights Series remains fair and unbiased, identifying and flagging any biased items as soon as DIF can be identified with sufficient statistical certainty. As this process is automated, the Insights Series can run far more DIF analyses than would be possible by a human administrator. The Insights Series will investigate evidence of DIF across all collected demographics, including (but not limited to) gender, nationality, ethnicity, education level, and employment status for every item within its three item banks. This constitutes millions of individual bias detection calculations during each analysis, generating vast amounts of information on potential item bias.

Automatic bias detection also allows for the detection of changes in the psychometric properties of items. Over time, an item may become more difficult for a specific demographic to answer correctly, but remain the same difficulty for another group. Any changes in the psychometric properties of an item specific to any group will be identified, regardless of when the divergence occurs. This ensures that all items are under constant review, and that bias can be detected regardless of when it starts to occur.

This automatic item bias detection function ensures that the Insights Series' level of fairness is maintained after launch and throughout ongoing item bank development. It also ensures that the Insights Series holds itself to a high standard of fairness, higher than most comparable psychometric assessments.

Flagged Item Procedure

To be flagged as a potentially biased item, several statistical requirements must be met. Firstly, the DIF contrast (the magnitude of the difference in difficulty between the reference and focal groups) must be large enough to indicate a moderate to large degree of DIF. If the DIF contrast meets or exceeds this threshold, additional conditions must be met, including sample size requirements, statistical significance tests, and standard error requirements. This ensures that any detected DIF is investigated thoroughly, and that flagged items are most likely genuine cases of item bias, and not mere statistical artefacts. Once a genuine case of DIF has been identified, a report is emailed to the item bank administrators, relaying statistical findings and advising the review and removal of the offending item(s). As the

Insights Series employs item banks, removing individual items does not interfere with testing or result in any disruption to clients.

Summary

Differential item functioning (DIF) analysis provides detailed insight into any potential item bias against specific demographics. Unlike tests of group differences using effect sizes, DIF aims to separate genuine differences in ability from bias, ensuring that biased items are identified even when effect sizes show no group differences. Insights bias detection conducts a vast number of DIF analyses automatically, searching for evidence of DIF on an ongoing basis and flagging potentially biased items for review and removal.

Ongoing item bank improvement

Although the Insights Series already comprises large item banks, ongoing item bank development will ensure that the item banks continually increase in size and improve in quality. The addition of new questions into the item banks will enhance overall reliability, security, and content quality, further improving the utility of the Insights Series.

The Flexibility of Item Banking

Item banks are repositories of questions with known psychometric properties calibrated on a common scale. When combined with item response theory methods, item banking can be used in several ways to administer items to candidates. Items can be administered to candidates either at random, using linear on the fly (LOFT) testing, via CAT or with a fixed set of questions. Item banking allows for virtually any combination of item selection procedures, freeing assessments from relying solely on fixed question sets. Considerable research is required to create and calibrate item banks, ensuring that items exist on a common scale. Nevertheless, the benefits achieved by item banking far outweigh their costs, providing a level of flexibility far beyond what would otherwise be possible.

Expanding Item Banks

Expanding item banks by adding new items maximises the advantages of item banking and CAT. The larger the item bank, the lower the probability of individual candidates receiving the same questions in their tests, improving item security. Also, the larger the item bank, the broader the range of items at different difficulty levels, ensuring optimal item targeting to improve reliability and precision. Expanding item banks also allows question topics to remain current and relevant, ensuring that the assessments remain engaging.

The Insights Series' item banks continually undergo additional calibration research, with the aim of further expansion. New items are trialled alongside the existing item banks, allowing new items to be included. As calibration occurs on a common scale, the addition of new items within an item bank does not disrupt users of the existing item bank, instead new items can be flexibly added and removed without disruption to candidates or clients. Once calibration has been achieved, and new items meet specific item quality standards, these new items are included in the Insights Series' item banks, appearing in future assessments.

Removing Items from the Item Bank

Once calibration research has been completed, removing existing items is just as simple as adding them. Although item writing guidelines for the Insights Series ensure that question topics avoid referencing dates and times in the future tense, inevitably information regarding certain topics will change and information within questions may become outdated. In this event, questions can be retired from item banks. As with the addition of items, the removal of items from an item bank does not cause any disruption to candidates or clients.

Continual monitoring of the Insights Series' psychometric properties also allows for the removal of items whose parameters change over time. Excessive item drift (a systematic change in item difficulty over time) can be detected by comparing the item's current difficulty estimate with its original difficulty from initial calibration. If key psychometric properties of items have changed, items can easily be reviewed and / or removed, ensuring that the quality of the item pool improves both by adding new items and removing underperforming items.

Summary

Item banking provides a high level of flexibility when administering assessments, freeing assessments from requiring fixed sets of items. The item banking strategy employed by the Insights Series allows for the continual addition of newly calibrated items, facilitating the ongoing improvement of item banks. Current items can also be conveniently removed from item banks should items become outdated or their psychometric properties change

Creation of bespoke and customised assessments

The flexibility afforded by item banking permits the creation of customised and bespoke versions of the Insights Tests. Organisations can request customised or bespoke assessments with specific question topics, subjects or content areas unique to their organisation, enhancing face validity and item security.

Creating Subsets of Existing Item Banks

The verbal and numerical reasoning assessments in the Insights Series contain questions based on a range of topics. With sufficient numbers, items from specific topics can be isolated, creating a sub-item bank of specific questions with similar topics. If a subset item bank can be created, specific clients may request customised assessments based on pre-existing items from the full item bank. For example, if a client requires a numerical reasoning assessment containing only finance related questions, a separate item bank containing only financial questions may be created from the original item bank.

Provided the original item bank contains enough relevant questions, a client specific assessment can be created solely based on existing items within that original item bank. However, if the original item bank contains an insufficient number of relevant items, the creation of a new item bank requires items to be developed and included. Due to the benefits of employing large item banks, creation of additional items would be advised when creating separate item banks for customised assessments.

Creating New Assessments

Certain clients may have specific requirements for an assessment, and thus require a bespoke assessment focusing on a highly specific set of topics. Development of bespoke assessments alongside existing item banks simplifies the development process, as previous R&D carries over to the new assessment. This ensures that the new assessment benefits from the previous research conducted on the original item bank, rather than starting an entirely new research project. The resulting assessment then has the flexibility to remain inside the original item bank, or as a subset within the original item bank, forming a stand-alone assessment.

Developing new assessments using existing item banks can considerably reduce the test development time, allowing for faster delivery of bespoke assessments. Similarly, new bespoke assessments will benefit from the previous research conducted on the original item bank, creating more robust assessments than would otherwise be possible. This makes bespoke assessments with specific topics more accessible to employing organisations, benefiting that organisation in several ways.

Benefits of Bespoke and Customised Assessments

Higher Face Validity: Bespoke assessments designed with a specific role in mind can achieve higher levels of face validity than off-the-shelf assessments. When candidates are provided with question topics relevant to the role, candidates readily recognise the relevance and potential usefulness of the assessment in that specific context. Candidates who are unfamiliar with psychometric testing and related research may harbour scepticism towards the use of psychometric tests in employee selection. Bespoke measures may reassure these candidates, helping them to recognise the utility of ability tests in a selection and assessment setting.

Stakeholder Buy-In: In addition to ensuring candidate buy-in, bespoke assessments may increase the buy-in of additional stakeholders. General management may be more inclined to prioritise the use of face valid bespoke assessments over generic measures, maximising the utility of the assessments. It may also ensure that higher level management provide sufficient resources to test administrators by recognising the importance of aptitude testing.

Benefits to Item Security: Unlike generic, off-the-shelf assessments, newly created bespoke assessments are employer specific, with only that employer's candidates completing the assessment. The smaller the total number of candidates completing the assessment, the lower the overall item exposure and therefore the higher the item security. It also ensures that, in the case of candidates applying for roles at several different organisations, there is no risk of candidates completing the same assessment on behalf of another employer, further maximising test security.

Employer Branding: Using customised and bespoke assessments relays organisational commitment to recruitment. It signals to applicants that time and resources have been invested into designing an effective and engaging selection process, highlighting that recruitment is a priority to that organisation. This helps ensure that applicants retain a high opinion of the employing organisation throughout the selection process and beyond.

Summary

The Insights Series' item banking strategy allows for the development of customised and bespoke assessments for specific employers. Switching from generic to customised assessments can occur seamlessly, with the scores automatically equated between the two types of assessment. Customised and bespoke assessments can be created as entirely new assessments, or retain some items from the main Insights item banks depending on the employer's requirements.

Section 4: Development of the Insights Series

Overview

The Insights Series was developed over the course of four months, during which time the tests were calibrated using data provided by over 16,000 participants. The individual calibration samples collected for each of the Insights tests ranged from 4,612 to 6,123 participants. A breakdown of the development process is shown in the timeline below and repeated in Appendix C:

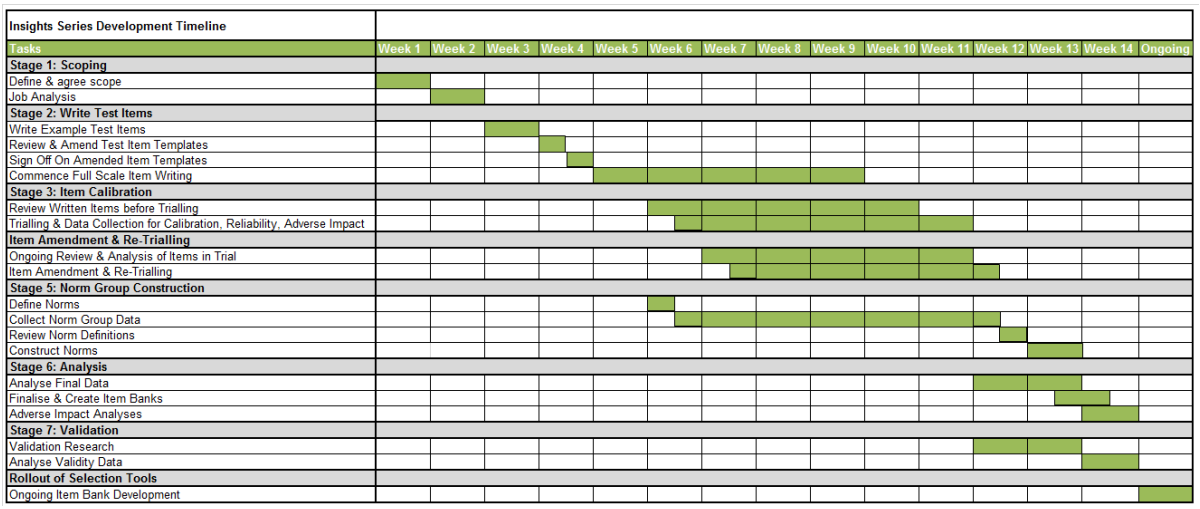

Figure 6: Development Timeline for The Insights Series

Item Writing and Reviewing

After an extensive review of the literature regarding item writing best practice for Rasch calibration, item writing specifications for each test type were produced and finalised. The item writing specifications were distributed to trained item writers, whose work was overseen by business psychologists at regular intervals. Each test item was subject to multiple checks, during which the style, grammar, spelling, clarity, reasoning, and question length was reviewed and amended. This was to ensure that all items were written in accordance with the provided specifications, prior to data collection and calibration.

Item Style and Structure

Each item met strict specifications for subject matter relevance, question length, image size, and clarity. They were also designed to be as concise and unambiguous as possible. Negatively-worded questions were avoided because they can be unfairly misleading, and items deemed too similar to one another were discarded in an attempt to avoid the issue of local dependence. Both the Insights Numerical Reasoning and the Insights Verbal Reasoning tests display multiple questions per text passage / infographic, with three corresponding questions per infographic and four corresponding questions per text passage, respectively. By doing so, this is considered to reduce the burden placed on the candidate, as they do not need to familiarise themselves with a new set of information for each question administered. Abstract Reasoning items were designed with an emphasis on clarity over

detail, so that candidates are not hindered by excessive detail when attempting to identify patterns. The Insights Inductive Reasoning test displays one question per image, as the multiple question format is not applicable to this type of test. All items were designed to require no greater than GCSE level calculation or reading proficiency, to remove advantages of prior knowledge and create a fairer testing experience.

Although psychometric tests do not require face validity, there are benefits associated with tests that have recognisable relevance to the application. Assessments that appear relevant to the role can provide the advantage of candidate perceived credibility. This in turn could encourage candidates to take the tests seriously and accept the outcome associated with their test performance. For this reason, the Insights Series was designed to be commercially relevant, without requiring candidates to have prior knowledge regarding any subject matter.

Trialling Procedure

Approximately 250 items were trialled for each of the three tests (Numerical, Verbal and Inductive Reasoning). During item calibration, items were trialled online and administered randomly to a large sample of potential candidates. Participants were directed to the test trials via a network of practice test websites, meaning that calibration data were provided by participants who likely sought to prepare themselves for upcoming assessments and thus were more likely to be intrinsically motivated to provide honest responses. This was reflected by the large proportion of participants who completed the test trials fully. Participants did not have the option to skip questions or progress further without providing responses. A cooldown period which set a minimum response time was applied to each item, which was designed to deter test completion in those who did not take the test trial seriously. These precautions were taken in attempts to safeguard against dishonest participants and ensure that only high-quality data were collected. Demographic questionnaires were also included at the trialling stage, to allow for norm group construction once item trialling had been completed.

Data collection for the purpose of item calibration took place over the course of approximately six weeks, during which time ongoing statistical analyses were performed asynchronously during data collection. If necessary, items were amended and re-trialled based on the findings of the ongoing statistical analyses. Extreme scores and responses from people who did not complete all items within a test trial, were removed from the final data sets and thus were not included in analyses.

Calibration Statistics

To ensure stability in item difficulty calibrations, the modelled standard error must be sufficiently low, which requires data collection from large calibration samples. When calibrating items using the Rasch model, stable estimates can be achieved with as little as 30 participants per item. However, high stakes tests (as used in candidate selection) require much larger sample sizes in order to reach a satisfactory confidence interval. The recommended sample size required to calibrate items within a high stakes test, at the 99%

confidence interval, is n=250 per test item (Linacre, 1994). The minimum calibration size for individual test items within each of the Insights tests were as follows:

Insights Verbal (*n=1100*)
Insights Numerical (*n=448*)
Insights Inductive (*n=444*)

Item Selection and Quality Control

Items were calibrated using the Rasch model and Infit Mean-Square (INFIT) statistics were chosen as the primary quality control statistic (Wright, Linacre, Gustafson & Martin-Lof, 1994), as well as point measure correlations (PTMA). The INFIT statistic was selected as it is information-weighted (Linacre, 2002), and the adaptive testing protocol of the aptitude tests ensures item targeting (Linacre, 2006a). Point measure correlations were checked to ensure that each item was consistent with the construct being measured, as indicated by a positive figure. After calibration, all items with an INFIT statistic in excess of 1.3 were rejected from the final item bank, in line with item selection guidelines for multiple-choice format tests (Wright, Linacre, Gustafson & Martin-Lof, 1994). Items with a negative point measure correlation were also removed from the item bank.

From launch, each item bank will increase in size on an ongoing basis, as new items are trialled and included. This is part of an ongoing development initiative to ensure that the quality of each item bank is constantly improving and that each test remains relevant.

Item Targeting

Each item bank contains questions with a broad range of difficulties, so that candidates of all abilities will have suitable items available from the item bank. The dispersion of item difficulties for each item bank was analysed and interpretations of Wright maps suggested that the range of item difficulties matched that of the range of person abilities. This suggests that the items available are suited to providing an appropriate ability measure. The number of items available at a certain difficulty varies in accordance with the normal distribution curve. This means that a large proportion of items within each item bank will be an average level of difficulty, as these will be administered most frequently. Items commensurate with extreme levels of ability, i.e. extremely difficult or extremely easy, are required in smaller volumes than those of average difficulty, as few of these are likely to be administered.

Section 5: Issues of Fit and the Assumptions of the Rasch Model

Overview

The Rasch model has a set of requirements that must be reasonably met in order to produce high quality measures of the target variable. Unlike other statistical methods where the underlying assumptions of the model must not be violated, the Rasch model has ideals that must be sufficiently approximated. Instead of determining the suitability of the data to a specific method prior to analysis, the Rasch model incorporates this into the analysis itself. During which, both items and persons are assessed for fit to the model and its underlying ideals. If the quality of data produced from items or persons does not sufficiently fit the Rasch model, it is removed from the dataset. Simply put, the data are made to fit the model, as opposed to selecting a statistical method based on the nature of the data available.

The main issues in terms of "fit" to the Rasch model are equal item discrimination and absence of guessing. The main assumptions of the Rasch model and item response theory (IRT) are local independence, invariance and unidimensionality. To constitute fit to the Rasch model and satisfy its assumptions, data are not expected to fulfil these requirements exactly, but must do so within reason.

Issues of Fit

Item Discrimination – The Rasch model holds the ideal that all items within an item banked test should have equal discrimination. Discrimination refers to the extent to which successful performance on an item will relate to successful performance on the overall test. Items with negative or zero discrimination in terms of the measured variable are redundant, if not detrimental, as they do not relate to the latent trait. Positively discriminating items provide productive measures, provided that the discrimination is not so high it suggests mere repetition of other items. High positive item discrimination due to inclusion of items that are extremely similar to one another is known as the attenuation paradox, for which data should be checked.

The discrimination of an item can be shown on an Item Characteristic Curve (ICC), which is an S-shaped curve that illustrates an item's properties. The difficulty of an item corresponds to the point on the curve at which there is a 50% chance (p=0.5) of answering the question successfully. The discrimination of the item is the slope of the curve at p=0.5.

During Rasch analysis, "fit" statistics are calculated to determine how accurately and predictably the data fit the model. Mean-square fit statistics provide an indication of the amount of measurement distortion or "noise" in the data and are commonly referred to in terms of under fit and over fit, using the Infit Mean-Square (INFIT) statistic. For data to fit the Rasch model perfectly, the INFIT of an item would be 1.0, although this rarely occurs with real world data. Items that display INFIT statistics close to 1.0 demonstrate very little measurement distortion and can be included in the item bank, provided the INFIT does not deviate greatly from this figure. In terms of tests that adopt a multiple-choice format, acceptable magnitudes of INFIT statistics are recommended to range between 0.7 and 1.3, with an excess of 1.3 indicating under fit and less than 0.7 indicating over fit to the model

(Linacre, 2002). All items included in the Insights Series had INFIT statistics within the range 0.7 – 1.3 indicating sufficiently equal discrimination across items. Items that did not produce an INFIT statistic within the accepted boundary, were rejected during item bank construction.

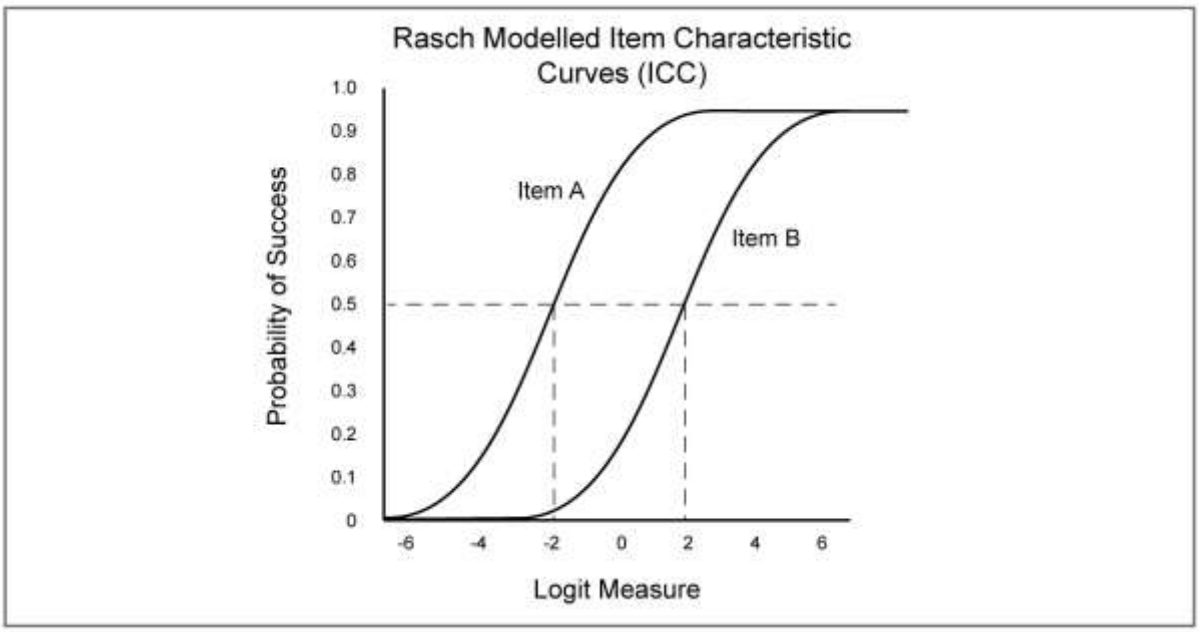

Figure 7: Rasch Modelled ICC's of Two Items with Different Difficulty (Logit) Measures

Guessing - As with any Multiple Choice Questionnaire (MCQ) or test that employs the MCQ response format, guessing could pose a threat to measurement accuracy. This is especially true during item calibration, as successful responses on ability tests due to guessing are not considered adequate measures of the latent trait. However, there is some evidence that during high stakes testing, candidates are more likely to adopt guessing strategies when the items they receive are too difficult for them (Waller, 1974). Computer adaptive test (CAT) protocols, as used in the Insights Series, are designed to administer test items with difficulties that are somewhat commensurate to the candidate's ability, meaning that candidates should only ever receive test items that they have a good chance of answering correctly. Generally, Rasch calibrated CATs administer items that give candidates a 50% probability of success, as this is considered optimal item targeting (Linacre, 2006a). However, the Insights Series is designed to administer items that candidates have an approximately 70% chance to answer correctly. With this in mind, the difficulty of items administered to candidates should dissuade guessing.

Unlike the 3 and 4 Parameter Logistic (PL) IRT models that parameterise guessing, altering item difficulties and person abilities in accordance, the Rasch model does not account for item guessability and instead models this as misfit. Rasch purists see this as an advantage, due to the discrepancy between informed and random guessing. It is suggested that some guessing behaviour is based on informed decision making, which is likely to require a level of ability and thus contributes towards calibration and measurement (Smith, 1993). This differs from random guessing, in which success is due to chance and thus does not contribute to calibration or measurement. The Rasch model allows the analyst to assess the quality of the data, whereas it could be argued that parameterising all guessing behaviour could automatically remove data that may be valuable.

As a precautionary measure against guessing, a number of statistics can be used during Rasch analysis, such as the Infit Mean-Square (INFIT) which indicates whether items or persons behave as the Rasch model would expect. For example, a person who answers items of a certain difficulty incorrectly, would not be expected to perform successfully when presented with an item of greater difficulty. This could suggest that the unexpected success on the high difficulty item was due to random guessing. All items included in the Insights Series had an INFIT between 0.7 and 1.3, which is considered to indicate very little measurement distortion and demonstrates the recommended boundary for tests with a multiple-choice format (Linacre, 2002). Further reference is made to the INFIT statistic in the segment titled "*Item Discrimination*" (see section 5).

Observing lower asymptotes and adopting a "CUTLO=" method to eliminate off-target responses, also offers protection against guessing. When using CUTLO, the analyst can set a minimum person ability measure, based on observation of the lower asymptotes, in order to disregard responses from persons who display evidence of random guessing behaviour. Data were assessed during item bank construction and it was not deemed necessary to employ a CUTLO method. Items that showed evidence of encouraging random guessing, were rejected from inclusion in the Insights Series.

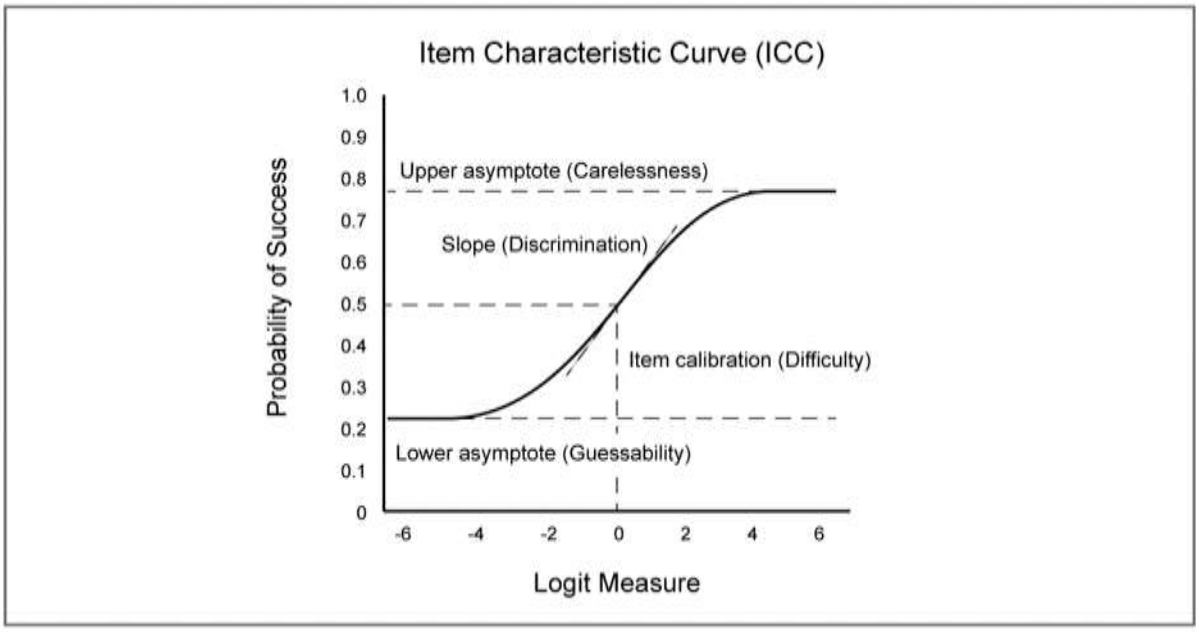

Figure 8: An ICC Demonstrating the Locations of Possible Parameters

Carelessness - Just as candidate guessing could pose a threat to measurement accuracy, the same is thought of candidate carelessness. This could occur during item calibration if respondents do not take the test trial seriously or are not sufficiently invested in their performance. If misleading or "trick" questions are present, carelessness may be more evident as respondents may be more likely to fall for cognitive pitfalls if they do not give the test trial their full attention. Both carelessness and trick questions could distort calibration measures and for this reason, data should be checked for their occurrence.

If there is sufficient evidence that respondents demonstrate carelessness, the "CUTHI=" method can be adopted. CUTHI removes responses from all persons who are *CUTHI= Logits* above the difficulty of the item, meaning that high ability people who are unexpectedly unsuccessful on low difficulty items, are not included in item calibration. This is under the

assumption that their incorrect responses were due to carelessness and would otherwise distort the data. Determining the CUTHI boundary can be informed by referring the upper asymptotes, unlike CUTLO in which the lower asymptotes are referred to. Data were assessed during item bank construction and it was not deemed necessary to employ a CUTHI method. If candidate responses showed evidence of carelessness, their responses were removed from the calibration data.

Possible trick questions were investigated based on fit statistics and distractor analyses. Items that exhibited properties of trick questions or appeared to be misleading, were rejected from inclusion in the Insights Series.

Assumptions of IRT and the Rasch Model

Local Independence of Items: The Rasch model assumes that all items are independent of one another, which means that the answer to a question should not be informed by determining the right or wrong answer to a previous question. If the process of selecting or rejecting one of the possible responses, contributes towards discerning a right or wrong answer to a distinct question within the same test, this is considered local item dependence (LID). Local dependence can be problematic as it can artificially reduce standard errors of estimates (SEE) and inflate reliability to the detriment of a test, as the questions resemble one another too closely. Local dependence also effects the dimensionality of a test as it introduces bias to measurement, possibly constituting a subsidiary dimension. Local dependence can be assessed by looking at residual correlations between items. Once the contribution of the latent trait has been removed there should not be significant correlations between items, as this would suggest the residuals share an underlying dimension that differs to the main Rasch dimension. Generally speaking, local dependence should only be considered a practical concern once residual correlations between items reach a minimum of r=0.7 (Linacre, 2006b). This threshold indicates that items must have at least 49% of the variance in their residuals in common, to constitute a problem with local dependency.

Statistical analyses to determine whether local item dependence was an issue across each test in the Insights Series produced correlations that ranged from -.10 to .17. All of which were notably lower than the recommended threshold of 0.7. Based on these findings, the Insights Series can be considered to satisfy the assumption of local independence of items.

Unidimensionality: Unidimensionality refers to the nature of the latent variable measured and suggests that a singular construct should be present. Occasionally, there can be evidence of multiple sub-facets which are expected to comprise the overall construct and in such cases, this should not be considered problematic (Linacre, 2009). However, if analysis suggests that multiple constructs are present which are likely to bias overall measurement, this is a problem and the appropriate action must be taken (Linacre, 2009). The Rasch model is a unidimensional measurement model, which means that data are always analysed under the assumption that they are unidimensional (Linacre, 2011). It is then up to the test developers to refer to the reported statistics in order to determine whether the data sufficiently match the unidimensional framework constructed during Rasch analysis, this is to say that the assumption of unidimensionality has been adequately met.

Principle-Components analysis (PCA) of residuals is a useful method to determine the dimensionality of a variable. It differs from traditional factor analysis as it does not aim to construct variables, but to explain variance. PCA of residuals seeks to falsify the hypothesis that residuals are random noise and determine whether patterns among the residuals can be explained by a common factor, or subsidiary dimension. During PCA of residuals, the eigenvalue should be less than 2.0, which is equivalent to the strength of two items within the test (Linacre, 2015). If the total variance explained by residuals is less than 4%, this can be attributed to random noise (Linacre, 2015).

PCA of residuals for each test in the Insights series indicated that the assumption of unidimensionality was met in each case. All residual eigenvalues were below 2.0 as recommended and the percentage of variance explained by residuals ranged from 0.6% - 1.2% across all three Insights tests.

Invariance - In Rasch analysis, items are assigned a difficulty measure and persons are assigned an ability measure. Item invariance is the assumption that item difficulties remain stable with respect to irrelevant aspects of the candidates, such as demographic differences and specific group membership. However, differences that are related to the assessment, such as use of a calculator during a numerical reasoning test, is expected to produce some variance in item difficulties. Item invariance can be assessed in terms of Differential Item Functioning (DIF), which looks at whether items appear to function differently depending on a candidates' group membership, despite having equal person abilities. As an example, a man and a woman with equal ability could perform differently on the same item. If this were to happen consistently, it may suggest that the item is biased towards a certain group. This is of fundamental importance in terms of test bias, as it is essential that all candidates have equal opportunity to perform well irrespective of group membership, and not be hindered by extraneous variables such as bias. As with adverse impact analysis, DIF analysis can be used to assess whether items perform at the same difficulty level across protected groups, i.e., gender, age, and ethnicity. During DIF analysis, no item met the standard requirements to constitute evidence of differential item functioning across protected groups. Therefore, items within the Insights Series can be considered to sufficiently satisfy the assumption of item invariance.

Section 6: Reliability

Overview

This section of the manual explains the importance of reliability and reports the findings of investigations into the reliability of the Insights Series.

Introduction to Reliability

In the context of psychometric assessments, reliability refers to the accuracy, precision, and replicability of assessment scores. Inevitably, any score on an assessment can never be a 100% accurate measure of someone, and thus observed scores are always estimates of ability based on test performance. The level of reliability provides an estimate of that assessment's level of precision, with higher levels of reliability indicating greater measurement precision than lower levels of reliability. Numerous factors may influence the reliability of an assessment, most notably the quality and quantity of questions within the assessment. Therefore, ensuring that assessments are of sufficient length and that their questions are of demonstrably high quality is essential in ensuring necessary test reliability. Reliability is unescapably connected to validity, as sufficient reliability is a prerequisite of test validity. If scores are inaccurate, inconsistent or imprecise, attempts to validate the assessment will inevitably fail. Although a high level of reliability is not *prima facie* evidence of validity, sufficient unreliability will prevent validity, rendering the assessment's scores useless in practice. However, if an assessment is demonstrably valid, increasing its reliability will further increase its validity, improving the utility of the assessment as a predictive tool.

Reliability in Classical Test Theory (CTT) vs. Item Response Theory (IRT)

Although reliability in IRT differs from CTT conceptions of reliability, both approaches show similarities. In both cases, reliability statistics range from 0.00-1.00, with .70 considered the minimum acceptable level of reliability for testing. As a result, both CTT and IRT reliability provide estimates of precision and accuracy within the same frame of reference. IRT reliability statistics are therefore roughly analogous to CTT reliability statistics, albeit with a number of notable differences.

In CTT, the quality and number of questions in the assessment primarily dictate the reliability of the assessment, with assessments comprising a large number of high quality questions displaying the highest levels of reliability. This is also true in IRT, but additionally the item difficulty targeting significantly influences the reliability of the assessment, as highlighted by the higher levels of reliability displayed in CATs. In IRT based assessments, the better the item targeting, the more information provided by the administered questions and thus the higher the reliability, independent of the number or quality of the questions.

Another key difference between IRT reliability and CTT reliability is that in IRT, reliability estimates can be person specific, rather than just test specific. CTT reliability estimates, such as Cronbach's alpha are generated at the test level, indicating an average level of reliability across all persons measured. IRT reliabilities however, can provide a test level and a person level estimate of reliability, displaying how reliable the assessment was in the case

of each individual tested. This provides insight into a test's level of reliability at different candidate ability levels, which holds significant practical implications.

In CTT based assessments, the measurement precision of top and bottom performing candidates tends to be lower, as fixed form tests tend to contain large numbers of moderate difficulty questions, with no level of adaptivity to candidate's abilities. This results in a high level of measurement precision at the middle ability range, with scores for high and low performers showing lower levels of precision. In CATs, the reliability estimate for each candidate is roughly equiprecise, showing comparable levels of reliability regardless of the candidate's ability. Moreover, as the number of items administered during CAT is flexible, it allows test developers to choose a minimum level of reliability for each candidate, guaranteeing sufficient reliability at every level.

Cronbach's Alpha vs. Rasch Reliability

Cronbach's alpha is the most commonly used estimate of internal consistency reliability in CTT based assessment. High Cronbach's alpha coefficients are typically seen in long assessments with a large number of high quality items, as measured by their item-total correlations. Rasch person reliability is roughly analogous to internal consistency as measured by Cronbach's alpha, but with a few important distinctions. Rasch reliability is a function of the standard error (SE) estimate of each person's score and the spread of person abilities in the sample. This SE, like Cronbach's alpha, is dictated by the assessment's length, but also by the difficulty targeting of the assessment, with CATs showing lower SEs than fixed form or randomised assessments.

Another key difference between Cronbach's alpha and Rasch reliability is that Rasch reliabilities can be estimated in the presence of missing data, permitting their use in item banked and CAT assessments. Cronbach's alpha requires a complete data set, with missing data severely impacting the accuracy of the reliability estimate. Rasch reliability however, is unaffected by missing data, providing an equally accurate estimate of reliability independent of missing data. This is especially useful in item banked assessments, as each candidate is likely to receive a small subset of the items within the bank.

There is also a practical difference between Cronbach's Alpha and Rasch person reliability, in terms of how conservative their reliability estimates are. Cronbach's alpha tends to overestimate reliability, inflating the coefficient compared to other forms of reliability (Linacre, 1997). Rasch person reliability however, tends to underestimate reliability, providing a more conservative estimate of reliability. Extra-conservative measures of reliability can be provided by reporting "real" Rasch reliability statistics, instead of "model" Rasch reliability coefficients, which inflate SE estimates based on person fit statistics (Wright, 1996). The difference between Cronbach's Alpha and Rasch person reliability estimates from the same dataset presented in Figure 9.

SUMMARY OF 113 MEASURED (NON-EXTREME) Person

	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	11.4	18.0	.80	.61	1.00	.0	1.00	.1
P.SD	3.4	.0	1.19	.12	.20	.8	.48	.8
S.SD	3.5	.0	1.19	.13	.20	.8	.48	.8
MAX.	17.0	18.0	3.32	1.06	1.45	1.9	4.29	2.6
MIN.	2.0	18.0	-2.52	.53	.57	-2.0	.30	-1.5

REAL RMSE	.65	TRUE SD	.99	SEPARATION 1.52	Person RELIABILITY	.70
MODEL RMSE	.63	TRUE SD	1.01	SEPARATION 1.61	Person RELIABILITY	.72
S.E. OF Person MEAN = .11						

CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .77

Figure 9: Rasch Reliability Data Output

In the case of the example dataset shown above, a Cronbach's Alpha of .77 was analogous to a "real" Rasch person reliability of .70 and a "model" Rasch person reliability of .72. Although the difference between the two coefficients is test specific rather than uniform, Rasch reliability of the same assessment will typically display a lower number than the Cronbach's alpha from that same assessment. Therefore, setting a minimum standard of .70 for Rasch reliability implies a minimum Cronbach's alpha equivalent equal to or greater than .70.

Rasch Person and Item Reliability

The Rasch model provides two separate forms of reliability, person reliability and item reliability. Person reliability is analogous to traditional notions of test reliability, indicating a level of precision when measuring the ability of the persons being measured. Rasch person reliability can be interpreted in the same frame of reference as Cronbach's Alpha, with a minimum of .70 as an acceptable level of test reliability.

Factors which influence Rasch person reliability include:

1. Sample ability variance: Wider ability range = higher person reliability.
2. Length of test: Longer test = higher person reliability
3. Sample-item targeting: Better targeting = higher person reliability

Rasch person reliability is largely independent of total sample size, and can be calculated on a per-individual and per-test basis. Through the use of item banking and adaptive testing, the length of the test and the sample-item targeting are both manipulable variables, allowing test developers to control the reliability of the assessment, ensuring a predefined level of reliability to be achieved.

Rasch item reliability however, has no CTT equivalent, and is unique to the Rasch model. Rasch item reliability is concerned with the reliability of each item's difficulty estimate. A high level of Rasch item reliability implies accurate item difficulty calibration that is sufficiently free from error. Both Rasch item and person reliability statistics are calculated using the same formulas, however the minimum standard for Rasch item reliability is significantly larger than that of person reliability and CTT reliability, requiring hundreds of administrations per item to achieve.

Factors which influence Rasch item reliability include:

1. Item difficulty variance: Wide difficulty range = high item reliability
2. Person sample size: Large sample = high item reliability

Rasch item reliability is largely independent of total test length, and can be calculated on a per-item and per-test basis. When calibrating a large item bank, achieving sufficient item reliability can require many thousands of participants. However, this level of reliability is worth achieving because it ensures that item difficulty calibrations are accurate and precise; an essential requirement of item banking. When calibrating item banks using small samples, item difficulty calibration may provide inaccurate estimates, incorrectly representing the item's difficulty. This may result in easy items being erroneously rated as difficult, or vice versa, awarding candidates the wrong amount of credit during test administration. Therefore, ensuring sufficient item reliability is a primary concern during item bank development, and stringent item reliability requirements must be achieved before item trialling can conclude.

Reliability of the Insights Series

Both person and item reliability estimates have been generated for the Insights Series. Across all three assessments, the minimum standards for both item and person reliability were exceeded, ensuring high levels of precision. The reliability coefficients and related information are displayed below:

Person Reliability

During initial item calibration, a random selection of items was administered to each participant, ensuring that items were administered with roughly equal frequency. In the Insights Series, items are administered adaptively and with fewer items than during initial calibration. This means the Rasch person reliabilities obtained during calibration will differ from those expected during live assessments. As the item targeting procedure, the number of items administered, and the sample ability distribution of the populations being measured are known, Rasch person reliability can be estimated based on the available information.

The first step in estimating CAT reliability is estimating the standard error, which uses the formula below:

$$S.E. = \frac{1}{\sqrt{\sum_i (P(1-P))}}$$

*where **P** = item targeting and **i** = item administered*

Equation 5: Standard Error

Once the S.E. has been calculated, the reliability can be estimated using the following formula:

$$-\left[\left(\frac{S.E.}{SD}\right)^2 - 1\right]$$

*where **SD** = standard deviation of person abilities*

Equation 6: Reliability Estimate

These reliability estimates are population specific, and thus different levels of reliability can be expected across different norm groups. The estimated person reliabilities for the Insights Series are presented in Table 2.

All estimated reliabilities exceed the minimum required standard of .70, with the lowest estimated reliability at .81. This level of reliability ensures a high level of accuracy and precision when assessing candidates, regardless of the aptitude test used or the population being measured. Although these coefficients are already large, they are conservative values, as the method used to calculate the reliability underestimates reliability compared to Cronbach's alpha.

Norm Group	Insights Verbal *20 items*	Insights Numerical *15 items*	Insights Inductive *20 items*
Graduates, Professionals, Managers and Executives	0.85	0.87	0.81
Administrative, Operational, Apprentice and Non-graduate staff	0.84	0.88	0.83

Table 2: Insights Series Person Reliability Measures

The high level of reliability for each assessment across all norm groups, permits the use of the Insights Series either in combination or using each test as a standalone assessment tool. Similarly, when selecting cut-off scores or minimum pass-marks, assessors can apply standards based either on individual test scores, or on aggregated scores across multiple tests within the test series. Moreover, assessors can have confidence in the level of precision that the Insights Series offers, especially when making pass / fail decisions based on minimum scores.

Item Reliability

Rasch item reliability is based on the initial calibration samples used to generate item difficulty parameters. As the difficulty calibrations generated during item trialling are the difficulty parameters for the live assessment, the item reliabilities from initial calibration can be directly reported, rather than estimated. Item difficulty reliabilities for each Insights test are shown in Table 3 below:

Aptitude test	Sample Size	Difficulty SD	S.E.	Item Reliability
Insights Verbal	5998	1.03	.07	1.00
Insights Numerical	4612	.85	.11	.98
Insights Inductive	6123	.83	.10	.99

Table 3: Insights Series Item Reliability Measures

A general recommendation for Rasch item reliability is a minimum of .90. The observed item reliabilities range from .98-1.00, exceeding the minimum standard for Rasch item reliability. This high level of reliability ensures that item difficulty calibrations are accurate and precise, permitting their use in high stakes selection and assessment. It also ensures that item calibrations are highly stable, allowing these initial items to be used in common item linking, as part of ongoing trials for item bank expansion.

Summary

Reliability refers to the precision, accuracy and replicability of an assessment's scores. Both CTT and IRT methods are employed to estimate reliability, providing different, but analogous estimates. Rasch reliability, as used by the Insights Series, provides a more conservative estimate of reliability than CTT reliability coefficients such as Cronbach's alpha. High levels of reliability are estimated for the Insights Series, with even the most conservative estimates of reliability exceeding the required standards for both Rasch person and item reliability.

Section 7: Validity

Overview

This section of the manual explains the importance of validity and reports the findings of investigations into the validity of the Insights Series.

Introduction to Validity

The validity of a test is markedly important for multiple reasons. Primarily, validity is fundamental in order to establish whether a psychometric test is fit for its intended purpose. Validity tells us the extent to which a psychometric tool measures the target variable, and whether useful inferences can be made regarding the context in which this information will be used. Whereas reliability is a measure of how consistently and precisely the tool measures a variable, validity is a measure of how well a tool performs its intended purpose. Although reliability is a prerequisite for validity, reliability alone is not sufficient to determine the quality of a psychometric tool; it must demonstrate both validity and reliability. After all, no matter how precise or consistent the measurement, the tool is irrelevant if it measures a completely unrelated construct or does not suit its purpose. Similarly, if a psychometric tool does not produce reliable, accurate measurement, that tool cannot demonstrate validity. Validity can be contextual, as it is somewhat dependent on the requirement of the assessment. Evidence of validity when a tool is used in one circumstance does not necessarily equate to evidence of its validity in another.

Validity exists in a variety of forms, all of which contribute to the overall quality of a psychometric tool. Sufficient evidence regarding one or more of the following validity types must be assessed before a psychometric test can be considered valid. Depending on their intended purpose, some tests may require evidence of validity from multiple sources, whereas few sources of validity can be sufficient in some cases.

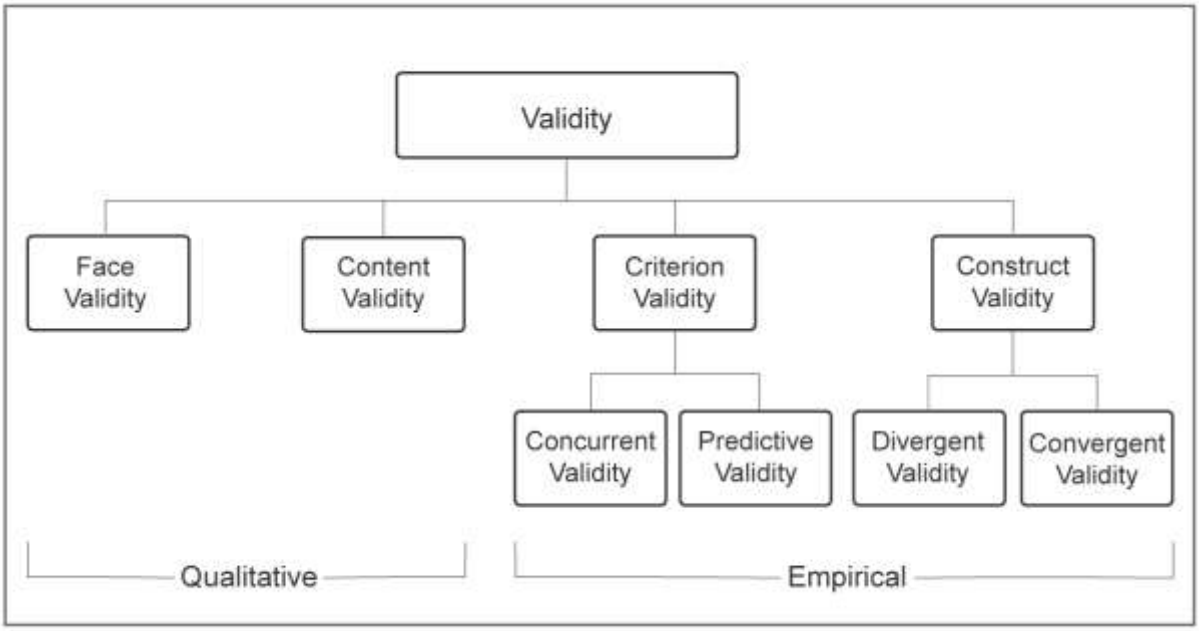

Figure 10: Validity and its Related Sub-Types

Construct Validity

Construct validity describes the extent to which a psychometric tool measures the variable or latent trait that it is intended to measure. In the case of the Insights Series this latent trait is cognitive ability, or specific sub-facets of cognitive ability if the tests are used independently of one another. Construct validity is fundamental to test development, as it provides evidence that the test successfully measures the target latent variable. For example, a test designed to measure verbal comprehension could unintentionally measure language proficiency instead, thus defeating its utility in the current context (unless a specific language proficiency is required in the role).

Evidence of construct validity requires in-depth empirical research and is usually determined via statistical analysis of the relationships between the test in development and existing measures of the intended variable. In this case, we would expect the Insights Series to exhibit positive significant correlations with established measures of cognitive ability, as they should measure the same underlying construct. Evidence of this kind of relationship between newly developed tests and pre-existing measures of similar psychological constructs, is known as convergent validity. Although it is arguably more important to determine what a test does measure, evidence of what the test does not measure can act as a form of construct validity in itself. For example, we would not expect to find large positive correlations between the Insights Series and measures of unrelated constructs. This type of construct validity is known as divergent validity and refers to establishing evidence of variance between the construct measured by the tests in development and constructs that we expect to be distinct, based on the available literature.

Evidence of construct validity can also be sourced from patterns of performance across specific populations. For example, we might expect engineers to obtain higher scores on a mechanical reasoning test than the general population, whereas we might also expect law students to outperform the general population when given a critical thinking task or verbal reasoning test. If patterns of performance replicate those of previous empirical studies regarding the target construct, then this can contribute towards identification of the variable measured.

Criterion Validity

Psychometric tests and assessments are generally used throughout recruitment processes to inform the decision-making process. For psychometric tools to do so, it is assumed that candidate performance based on such tools is related to a specific outcome. To establish criterion validity, test developers must conduct empirical research and provide evidence of a relationship between test scores and the external outcome of interest. In an occupational setting, this is usually job performance or aptitude for relevant skill acquisition. Such research requires a suitable criterion to be identified and thus fundamental understanding of the role, its requirements, and what constitutes its successful performance. Extensive knowledge of the variable being measured and how it relates to the target outcome helps to inform and underpin item writing guidelines. This in turn contributes towards the quality of the overall measure and its ability to fulfil its intended purpose. In summary, criterion-related

validity concerns a test or assessment's ability to indicate the likelihood of a specific real-world outcome. Given that this relates directly to the purpose of a psychometric tool, criterion-related validity, along with construct validity, could be considered the most important source of psychometric test validity.

Predictive validity and concurrent validity are sub-types of criterion validity and differ in relation to the nature of the criterion variable. Although both types of validity are concerned with establishing an association between test scores and an external outcome, concurrent validity involves acquiring evidence based on a criterion measure collected at the same point in time as the test scores, whereas predictive validity refers to a criterion measured subsequently to test performance. A criterion used to ascertain concurrent validity could be previous academic achievement, whereas a criterion to determine predictive validity could be a measure of job performance, having been successful during the recruitment process. Predictive validity could be the strongest evidence of a test's utility, as it refers to future job performance in relation to test performance. However, predictive validity is particularly difficult to establish as it requires tracking employee performance for a substantial period after the psychometric tools were used. This is challenging as the requirements of some roles can change rapidly, rendering some indicators of success invalid, as well as the data collected in relation to them.

Content Validity

Unlike construct validity and criterion validity which are assessed empirically, content validity is usually assessed qualitatively by subject matter experts during test development. Content validity refers to the extent to which the subject matter and content of a test or assessment are representative of the target construct, as well as the extent to which the target construct is adequately sampled. Content validity ought to be incorporated during the early stages of test development and should be informed by a thorough understanding and clear definition of the target construct. This should help to create a framework for item writing guidelines that ensures that the content reflects all aspects of the construct that the test is intended to measure.

Face Validity

As with content validity, evidence of face validity is usually determined on a qualitative basis. Although face validity is not a requirement in psychometric testing, it may improve candidate "buy in", which refers to whether candidates' perceive the test or assessment to be relevant in its current context. If candidates feel that they are being unfairly selected based on their performance on an unrelated test, they may feel less invested in the recruitment process or take the test less seriously. For instance, finance applicants may recognise the value of a numerical reasoning test, yet question the importance of a diagrammatic reasoning test without receiving further explanation. Candidates who recognise the relevance of an assessment may be more likely to accept decisions based its outcome. Face validity does not constitute utility or ensure that a test is suited to its purpose, instead it means that the test appears to be applicable to the recruitment procedure.

Validity Coefficients and Statistical Significance

The results of validity studies are expressed in terms of correlation coefficients, this refers to a figure that informs us whether there is a relationship between variables. Correlations range from -1 to +1, with a correlation coefficient of +1 indicating a perfect positive correlation. Correlations can be positive or negative, the direction of which signifies the direction of the relationship. After conducting extensive empirical research to establish evidence of construct and criterion validity, it is vital to interpret the findings correctly. Statistical significance should be considered, as this indicates the level of confidence we can have in the findings. A significance level of .05 indicates that there is only a 5% likelihood that the findings are due to chance, meaning that we can have 95% confidence that the results of statistical analysis represent a true finding, just as a significance level of .01 indicates 99% confidence in the findings and significance at .001 suggests a 99.9% confidence interval. Confidence intervals allow test developers to determine the degree of precision that can be expected from the findings. For example, a correlation of magnitude r =0.37 that is significant at the p = .01 level, indicates that the true correlation size lies between plus or minus 1% of 0.37. As with most statistical analysis, there will always be a likelihood that some findings are due to chance and confidence intervals help to determine the degree of that likelihood.

Statistically-significant correlations may inform us of a relationship between variables, but the magnitude and direction of these correlations is fundamental to determine the extent to which a test provides practical utility. The recommended boundaries for interpreting correlation sizes are 0.1, 0.3 and 0.5, with 0.1 - 0.29 considered “small”, 0.3 - 0.49 considered “moderate” and 0.5 and above considered “large”, regardless of direction (Cohen,1988). In terms of test utility, the size of correlations between test scores and the chosen construct or criterion variable are considered to indicate how effective the tests will be during recruitment, provided that the test is suited to the recruitment purpose. Tests that demonstrate significant correlations in the range 0.20 – 0.35 are considered to prove “quite useful”, whereas those with correlations to the construct or criterion variable of over 0.35, are considered to be highly effective (Murphy & Davidshofer, 1998).

Validity of the Insights Series

The validity evidence regarding the Insights Series is provided below. However it is desirable to undertake further research to establish the relationship between the tests and job performance. We encourage users to submit any relevant data that they collect so that we can build construct and criterion-related validity evidence for the Insights Series. For the design and analyses of these validation studies we are happy to advise and support users in this.

Construct Validity of the Insights Series

Construct Validity Study 1: Intercorrelations Between Insights Tests: Although ability tests measure a specific cognitive ability, all specific cognitive abilities are themselves partial measures of general cognitive ability. As a result we expect positive but not perfect correlations between performances on different ability tests. To evaluate the relationships

between the Insights tests, 995 participants completed questions from all three Insights tests. Scores for each test were then correlated with one another, and the results can be seen in Table 4 below:

Table 4: Intercorrelations between Insights Tests

Insights Tests	1	2	3
1. Insights Numerical			
2. Insights Verbal	0.55		
3. Insights Inductive	0.50	0.55	

All displayed correlations are significant at p<0.001.

The correlations between Insights tests range from 0.50 to 0.55, showing strong intercorrelations between all three of the Insights tests. All correlations have been corrected for reliability, but not restriction of range. This evidence suggests that the Insights tests measure a similar underlying psychological construct, as we would expect with partial measures of general cognitive ability.

Construct Validity Study 2: Correlations with Performance on Cognitive Reasoning Tests:

Evidence of construct validity was established by correlating the measures obtained by each adaptive test with those obtained using pre-existing measures of similar constructs. Alternate measures of the target construct were administered to each individual after they had responded to a 15-item fixed form version of one of the three Insights tests. Correlations were interpreted in alignment with the magnitudes proposed by Cohen (1988), in which 0.1 is considered small, 0.3 is considered moderate and 0.5 or above is considered to be large. Once they had completed all items in a fixed-form Insights test, candidates were randomly assigned one of three construct validity measures; the ICAR, General Cognitive Ability tests or the Test Partnership Aptitude Suite.

Construct Validity – Composite Cognitive Ability Tests

Candidates assigned to complete this construct validity measure were administered the following cognitive ability tests in succession of one another (see Table 5 for correlations between aptitude tests and cognitive ability tests):

Cognitive Reflections Test (CRT): The CRT measures an individual's ability to reflect on problems, and override obvious (but incorrect) responses (Frederick, 2005).

BAROCO Short-form (BST): A short version of the original 100 item syllogism-solving problems (Shikishima, Yamagata, Hiraishi, Sugimoto, Murayama & Ando; 2011).

Berlin Numeracy Test (BNT): The BNT is a measure of numerical reasoning, risk literacy and statistical numeracy (Cokely, Galesic, Schulz, Ghazal & Garcia-Retamero; 2012).

Decision Making Competence Questionnaire – Applying Decision Rules Sub-test (ADR): One sub-test of an original set of seven behavioural decision-making tasks (Bruine de bruin, Parker & Fischhoff; 2007).

	Correlations with Composite Cognitive Ability Tests		
	Insights Numerical (n= 378)	**Insights Verbal (n= 281)**	**Insights Inductive (n=217)**
Cognitive Reflections Test	0.55***	0.42***	0.49***
BAROCO Short-Form	0.44***	0.48***	0.54***
Berlin Numeracy Test	0.59***	0.56***	0.20*
Applying Decision Rules	0.61***	0.62***	0.43***
Combined Cognitive Tests	**0.63*** **	**0.61*** **	**0.47*** **

*Note: * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 5: Correlations between Insights Tests and Cognitive Reasoning Tests

All correlations have been corrected for reliability, but not restriction of range. The correlations between the Insights tests and the Composite Cognitive Ability Tests can be interpreted as ranging from moderate to very large (Cohen, 1988). This suggests that the Insights tests measure a similar underlying psychological construct to that measured by the selected cognitive ability tests and thus suggests evidence of construct validity for the Insights tests.

Construct Validity – Test Partnership Aptitude Suite

The Test Partnership aptitude suite is a selection of tests that when combined, provide a measure of general cognitive ability (Schwencke & Guy, 2015). There are four individual aptitude tests within the suite:

Critical thinking (CT): This test is comprised of logical syllogisms which require deductions to be made, based only upon the information provided.

Verbal Reasoning (VR): This test is a measure of both understanding a passage of information, as well as what conclusions or assumptions can be drawn from it.

Numerical Reasoning (NR): This assesses the ability to interpret and comprehend numerical information, in order to complete the required calculations.

Inductive Reasoning (IR): This assesses the ability to recognise patterns and sequences in order to determine the next item in the sequence.

Table 6 displays the correlations between each of the Insights Series tests and the Test Partnership Aptitude Suite.

	Correlations with Test Partnership Aptitude Suite		
	Insights Numerical (n= 350)	**Insights Verbal (n= 309)**	**Insights Inductive (n= 193)**
Verbal Reasoning (VR)	0.56***	0.69***	0.32**
Numerical Reasoning (NR)	0.66***	0.42***	0.32**
Inductive Reasoning (IR)	0.54***	0.45***	0.53***
Full Aptitude Suite	**0.65*****	**0.58*****	**0.50*****

*Note: * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 6: Correlations between Insights Tests and Test Partnership Tests

All correlations have been corrected for reliability, but not restriction of range. The correlations can be interpreted as ranging from moderate to very large (Cohen, 1988), which further supports the evidence of construct validity for the Insights tests, as they appear to tap into a shared underlying construct

Construct Validity - ICAR

The International Cognitive Ability Resource (ICAR) is an open source cognitive ability test (Condon & Revelle, 2014). The ICAR contains four subtests:

Verbal Reasoning (VR): The VR subtest measures a person's ability to work with words and sentences.

Letter Number Series (LNS): The LNS subtest measures a person's ability to identify patterns in alphanumerical data.

3D Rotation (3DR): The 3DR subtest measures a person's spatial reasoning ability by rotating three-dimensional shapes.

Matrix Reasoning (MR): The MR subtest measures a person's ability to solve problems and think logically.

The Full ICAR is a composite of all four subtests, which forms a measure of general cognitive ability. Table 7 below displays the correlations between each of the Insights tests and the ICAR (as well as its subtests).

	Correlations with ICAR and ICAR Subtests		
	Insights Numerical (n= 322)	**Insights Verbal (n= 224)**	**Insights Inductive (n= 183)**
ICAR Verbal Reasoning	0.71***	0.74***	0.48***
ICAR Letter Number Series	0.47***	0.25*	0.29**

ICAR 3D Rotation	0.34***	0.44***	0.37***
ICAR Matrix Reasoning	0.48***	0.39***	0.54***
Full ICAR	**0.58*****	**0.55*****	**0.52*** **

*Note: * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 7: Correlations between Insights Tests and the ICAR

All correlations have been corrected for reliability, but not restriction of range. The correlations between the Insights tests and the Full ICAR can be interpreted as large (Cohen, 1988). This suggests that the Insights tests measure a similar underlying psychological construct to that measured by the ICAR and thus suggests evidence of construct validity for the Insights tests.

Although correlations for the subsets within the ICAR have been reported, it is important to note that only the correlations between the Insights tests and the full ICAR are the most indicative, as the ICAR was intended to provide a measure of general cognitive ability for which the full scale is required.

Criterion-related Validity of the Insights Series

Criterion-related Validity Study: Academic Achievement: Cognitive ability tests are useful predictors of many important life outcomes, in particular educational and occupational performance. Therefore, as cognitive ability measures, the Insights Series should display this kind of predictive validity.

A study was undertaken to determine the predictive validity of the Insights Series with academic achievement, in particular GCSE results. Participants completed one of the three Insights tests along with a demographic questionnaire, in which participants provided their GCSE grades for the three compulsory GCSE subjects in the UK.

GCSE results were correlated with test scores from each of the three Insights tests, these correlations can be seen in Table 8 below:

Insights Tests	GCSE Grade			
	Maths Grade	**English Grade**	**Science Grade**	**Combined Grades**
Insights Numerical (n=348)	0.49***	0.34***	0.44***	**0.47*** **
Insights Verbal (n=328)	0.35***	0.31***	0.29***	**0.36*** **
Insights Inductive (n =145)	0.40***	0.24*	0.30**	**0.35*** **

*Note: * = Significant at p<0.05, ** = Significant at p<0.01, *** = Significant at p<0.001.*

Table 8: Correlations between Insights Tests and GCSE Results

Table 8 above shows statistically significant correlations between the Insights Series and GCSE results. These figures have been corrected for unreliability, but not restriction of

range. These results are in line with the general academic literature regarding ability testing and academic achievement.

Note: Although ability tests are strong predictors of academic achievement, this does not imply that academic achievement is a strong predictor of job performance. Research shows job performance to be weakly correlated with academic achievement, whereas ability tests are consistently shown to be the strongest single predictors of job performance.

Section 8: Group Differences

Overview

This section describes the research conducted to investigate group differences in scores across the Insights Series. Although large samples of participants were used as part of the calibration, validity, and group differences research, it is recommended that organisations also carry out local adverse impact analyses.

The Importance of Group Differences and Adverse Impact

Group differences in a testing context relates to the extent to which there are differences in test scores achieved by different groups. If there are differences between certain groups on a test, then using the test may lead to hiring rates that adversely impact individuals from specific groups, depending on the size of the difference and the cut-off used. If group differences can be explained by bias, group differences on scores may constitute unfair and unlawful discrimination. The most common demographic variables assessed in group difference analyses are gender, ethnic group, and age.

Group Differences and Bias

As mentioned in Section 3, group differences do not necessarily imply bias. Similarly, the absence of group differences does not necessarily imply the absence of bias. To determine the presence of pervasive bias, differential item functioning (DIF) analysis must be carried out. However, unlike DIF, group differences provide insight into possible effects on selection processes and hiring rates across people from specific demographics, regardless of the root cause of those differences.

Methods to Assess Group Differences

An initial assessment of group differences can be conducted by looking at the size of the difference between groups. A commonly used method is the Cohen's *d* statistic. Cohen's *d* indicates the size of the mean difference between groups, represented in standard deviations. When interpreting effect size, a *d* value of 0.5 or greater can be considered of practical note.

While this gives an indication of the size of the difference between groups, the practical effects of the test is also affected by how it is applied. A test should only be used in situations where job analysis has identified that the attributes measured by the test are important for effective performance in the job. This should ideally be supported by validity evidence which demonstrates a relationship between test scores and job performance.

In practical use, group differences in test performance can be assessed by monitoring the pass rates of different groups based on the test results. If there is a substantial difference between the relative pass rates of different groups, then the test may be discriminating. To determine if the difference is substantial, the Four-Fifths Rule is often applied as a useful 'rule of thumb'. According to this rule, the pass rate of the minority group should be at least four-fifths (4/5 or 80%) of the pass rate of the majority group. For example, if the number of White applicants passing the test was 100 out of 200 (i.e., a pass rate of 50%), then the pass rate for Black or Asian applicants should be at least four-fifths of this – a pass rate of 40% or more.

The pass rates for the test will depend on the specific cut-offs applied by the test user. It is recommended that test users monitor the relative pass rates of different groups on an ongoing basis to ensure that the test is not discriminating at the cut-off level being used. Note that pass rates of groups where the sample size is small (i.e., less than 100 people) should be interpreted with caution, as the pass rates may be unreliable.

Group Differences and the Insights Series

The following protected groups were investigated for evidence of average group differences and unfair discrimination:

- *Gender:* Participants that reported to be male were compared against participants reporting to be female.

- *Language:* Participants that reported their first language to be English were compared against the participants reporting not to be fluent in English.

- *Ethnicity:* Participants that reported being white were compared against those who reported to be BME participants.

- *Age*: Participants that reported being under the age of 50 were compared against the average scores of participants reporting to be over the age of 50.

Protected Group	Group Differences in terms of Cohen's *d*		
	Insights Numerical	Insights Verbal	Insights Inductive
Gender	0.09	-0.03	0.17
Age	0.35	0.01	0.47
Ethnicity	0.18	0.49	0.16
Language	0.16	0.65	-0.12

Table 9: Average score effect sizes across different groups

All effect sizes (excluding “Language” for verbal) can be interpreted as being small or non-existent (Cohen, 1988). However, for verbal reasoning, the effect size for “Language” can be interpreted as being moderate (Cohen, 1988). This is to be expected as the Insights Verbal test is written in English, and requires a working knowledge of the English language.

Clients should only use the Insights Verbal test if English language proficiency is required for the role, or selecting candidates in English speaking countries. Inductive reasoning tests however are considered “Culture Fair” (Cattell, 1940) and can be used regardless of English language proficiency.

Summary

Group differences across the four studies generally report small to negligible group differences between focal and reference groups. The only effect size greater than 0.5 was observed in the Insights Verbal Reasoning assessment when investigating English as a first language. This result is to be excepted, as English language proficiency is required to complete a verbal reasoning assessment in English. The Insights Verbal reasoning test should therefore be used as a selection tool when English language proficiency is required for the role, or all candidates are fluent in English.

References

Anderson, J. O. (1998). Does complex analysis (IRT) pay dividends in achievement testing? *Paper presented at the Measurement in Evaluation:* Current and Future Directions for the New Millenium, Banff, Canada

Bergstrom, B. A., Lunz, M. E., & Gershon, R. C. (1992). Altering the level of difficulty in computer adaptive testing. *Applied Measurement in Education*, *5*(2), 137-149.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, *78*(3), 387-409.

Bond, T., & Fox, C. M. (2015). Applying the Rasch model: Fundamental measurement in the human sciences. Routledge.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of personality and social psychology*, *92*(5), 938.

Cattell, R. B. (1940). A culture-free intelligence test. I. Journal of Educational Psychology, 31(3), 161.

Cohen, J. (1988). Statistical power analysis for the behavioural sciences. Hillside. NJ: Lawrence Earlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *112*(1), 155.

Cokely, E.T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making,* 7, 25-47.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence,* 43, 52-64.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*(1), 13-21.

DeMars, C. (2001). Group differences based on IRT scores: Does the model matter?. *Educational and Psychological Measurement*, *61*(1), 60-70.

Eggen, T. J., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30*(5), 379-393.

Etnier, J. L., Salazar, W., Landers, D. M., Petruzzello, S. J., Han, M., & Nowell, P. (1997). The influence of physical fitness and exercise upon cognitive functioning: a meta-analysis. *Journal of sport and Exercise Psychology*, *19*(3), 249-277.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25-42.

Gardner, H. (1983). Frames of Mind: Theories of Multiple Intelligences New York.

Gottfredson, L. S. (1997). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence*, *24*(1), 13-23.

Gottfredson, L. S. (2003). Dissecting practical intelligence theory: Its claims and evidence. *Intelligence*, *31*(4), 343-397.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological bulletin*, *96*(1), 72.

Linacre, J. M. (1994). Sample Size and Item Calibration Stability. *Rasch Measurement Transactions*, *7*(4), 328

Linacre, J. M. (1997). KR-20 / Cronbach Alpha or Rasch Person Reliability: Which Tells the "Truth"? *Rasch Measurement Transactions, 11*(3), 580-1

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. *Chae, S.-Kang, U.–Jeon, E.–Linacre, JM (eds.): Development of Computerised Middle School Achievement Tests, MESA Research Memorandum*, (69).

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16*(2), 878

Linacre, J. (2006a). Computer-adaptive tests (CAT), standard errors and stopping rules. *Rasch Measurement Transactions*, *20*(2), 1062.

Linacre, J. M. (2006b, July 11). *Largest residual correlations for items. Retrieved from* http://www.winsteps.com

Linacre J.M. (2009) Unidimensional Models in a Multidimensional World, *Rasch Measurement Transactions*, 23:2, 1209

Linacre, J. M. (2011). Rasch measures and unidimensionality. *Rasch Measurement Transactions, 24*(4), 1310.

Linacre, J. M. (2015, March 15). Dimensionality investigation - an example. Retrieved from http://www.winsteps.com

Mount, M. K., Witt, L. A., & Barrick, M. R. (2000). Incremental validity of empirically keyed biodata scales over GMA and the five factor personality constructs*. Personnel Psychology*, 53, 299-323.

Murphy, K. R., & Davidshofer, C. O. (1998). Using and interpreting information about test reliability. Psychological testing: Principles and applications, 127-145.

Pelton, T. (2002). Where are the limits to the Rasch advantage. In *International Objective Measurement Workshop, New Orleans*.

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*, *124*(2), 262.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human performance*, *15*(1-2), 187-210.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: occupational attainment and job performance. *Journal of personality and social psychology*, *86*(1), 162.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. *Contemporary intellectual assessment: Theories, tests, and*, Issues, (3rd), 99-144.

Schwencke, B., Guy, L. (2015). Test Partnership's Adaptive Suite: Psychometric Properties and Validity, London: UK.

Shikishima, C., Yamagata, S., Hiraishi, K., Sugimoto, Y., Murayama, K., & Ando, J. (2011). A simple syllogism-solving test: Empirical findings and implications for g research. *Intelligence*, *39*(2), 89-99.

Smith, R. (1993) Guessing and the Rasch Model. Rasch *Measurement Transactions*, *6*(4), 262-3

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. CUP Archive.

Tang, K. L., Way, W. D., & Carey, P. A. (1993). The Effect of Small Calibration Sample Sizes on TOEFL IRT-Based Equating. *ETS Research Report Series*, *1993*(2), i-38.

Tonidandel, S., Quiñones, M. A., & Adams, A. A. (2002). Computer-adaptive testing: The impact of test characteristics on perceived performance and test takers' reactions. *Journal of Applied Psychology*, *87*(2), 320.

Waller, M. I. (1974). Removing the effects of random guessing from latent trait ability estimates. *ETS Research Report Series*, *1974*(1).

Waterhouse, L. (2006). Multiple intelligences, the Mozart effect, and emotional intelligence: A critical review. *Educational Psychologist*, *41*(4), 207-225.

Wright, B. D. (1989a). Dichotomous Rasch model derived from specific objectivity. *Rasch Measurement Transactions, 1*(1), 5-6

Wright, B. D. (1989b). Dichotomous Rasch Model derived from Counting Right Answers: Raw Scores as Sufficient. *Rasch Measurement Transactions.* 3:2 p.62

Wright, B. D. (1996). Reliability and separation. *Rasch Measurement Transactions, 9*(4), 472

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch measurement transactions, 8*(3), 370.

Van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

Appendix A: Psychometric Scales and Scores

Mean: The mean is an average of a sample of scores. It is calculated by adding all scores within the specified sample together and dividing the outcome by the number of observations included. The mean is one of the most commonly used measures of central tendency.

Standard Deviation: The standard deviation quantifies the spread of a set of data and indicates whether the values within a sample are close together or spread out. The standard deviation considers each value within a data set and its distance from the mean.

Normal Distribution: A normal distribution refers to a particular distribution of data, where the majority of observations lie at the mid-point and trail off towards the ends or "tails" of the distribution. The shape created by a normally distributed spread of data is often described as a "bell curve", as the standard normal distribution is symmetrical and resembles a bell shape.

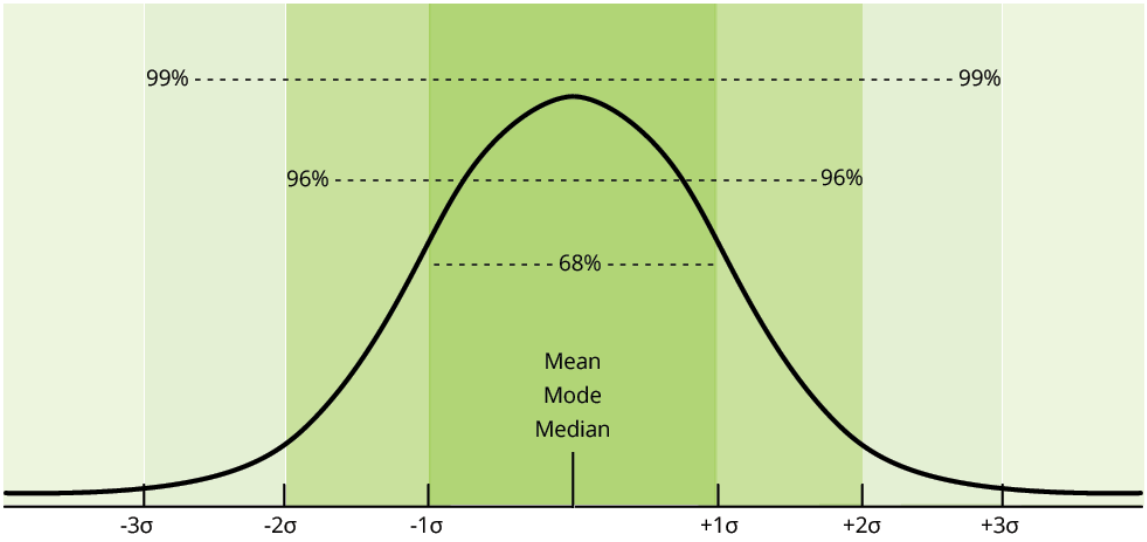

Figure A1: Normal Distribution Curve

The normal distribution is extremely useful when interpreting test scores relating to psychological constructs. For example, elements of personality and ability are assumed to be normally distributed among the general population. This means that most people tend to score within the mid-section or "average range" for these constructs, with extreme scores observed far less often, making the normal distribution particularly suited for use with psychometric tests.

The mean and standard deviation of the normal distribution can be used to determine the proportion of the population who scored better or worse than any given test score. This is because a standard normal distribution has specific statistical properties, as described below:

- The mean, median and mode all fall at the centre point of a normal distribution.

- 68% of scores fall within one standard deviation of the mean.

- 96% of scores fall within two standard deviations of the mean.

- 99% of scores fall within three standard deviations of the mean.

Norm Groups: A norm group is a sample of individuals who represent a specific population. This allows for comparisons between a candidate's score and the average score within the norm group, thus benchmarking their score against a representative population. Norm groups can be made for any demographic and require a minimum sample size of 150 people.

Percentiles: A percentile score together with the norm group (or 'comparison group') give context to a candidate's raw score. A percentile is a converted score that ranks where the candidate's raw score sits relative to the results achieved by people in the norm group. A percentile is defined as: *A value indicating the proportion of the norm group who scored less than the test taker.* Therefore, if a candidate scores in the 50th percentile, this means that the candidate's score is higher than 50% of the scores achieved by people in the norm group. If a candidate scores in the 90th percentile, they have scored higher than 90% of the norm group, putting them in the top 10%.

Percentile scores are not interval measures, which means the associated difference between them does not increase or decrease in equal increments, percentile scores simply rank the candidate's performance relative to the norm group. For example, a gap of five percentiles indicates a much larger difference between two candidates scoring within the 20^{th} and 25^{th} percentile, than the 50^{th} and 55^{th} percentile.

Standardised Scores: Standardised scores are raw scores converted to standardised scales, so that meaningful comparisons can be made on a common level. Standardised scales represent individual test scores in terms of their distance from the population mean, as represented by standard deviation units. Standardised scores such as Z Scores, T scores and Sten, differ from percentile scores as they are plotted on scales that have equal intervals between each value.

Z Score: Z scores have a mean of 0 and a standard deviation of 1. They are arguably the most important standardised values in this case, as it is from Z scores that raw scores can be transformed into other standardised scales. The formula for which is shown below:

$$\textbf{Standardised Score Scale} = (\mathbf{Z} \times \mathbf{SD}) + \overline{\mathbf{X}}$$

Z = *Z Score* **SD** = *Scale Standard Deviation* $\overline{\mathbf{X}}$ = *Scale Mean*

Equation A1: Conversion of Z Scores to Alternative Standardised Scales

T Score: A standardised value converted from a Z Score. It is expressed on a standardised scale that has a mean of 50 and a standard deviation of 10.

Sten: A portmanteau of "standard ten", a Sten is a standardised value converted from a Z Score. It is expressed on a standardised scale ranging from 1 – 10, that has a mean of 5.5 and a standard deviation of 2.

Appendix B: Norm Group Information

Insights Numerical

There are two norm groups currently available for Insights Numerical:

Norm	Description	Sample Size
1	Graduates, Professionals, Managers and Executives	2244
2	Administrative, Operational, Apprentice and Non-graduate staff	629

Information on each of these norm groups is presented in this section.

Norm 1: Graduates, Professionals, Managers and Executives

Norm 1 comprises 2244 Graduates, Professionals, Managers and Executives. The biographical data for this sample is shown below:

Gender

54.0% of the sample (n=1211) was male and 41.1% (n=923) female. 4.9% of the sample (n=110) did not indicate his or her gender.

Nationality

Table B1.1.1 below shows the sample broken down by Country of Nationality. 57.5% of the sample indicated that they were UK Nationals.

Country of Nationality	Frequency (N)	Percent of Sample
Angola	2	0.1
Argentina	1	0.0
Asia/Pacific Region	2	0.1
Australia	66	2.9
Austria	2	0.1
Azerbaijan	5	0.2
Bangladesh	6	0.3
Belarus	2	0.1
Belgium	26	1.2
Botswana	1	0.0
Brazil	2	0.1
Brunei Darussalam	3	0.1
Canada	21	0.9
Cayman Islands	1	0.0

Chile	1	0.0
China	5	0.2
Colombia	2	0.1
Cyprus	2	0.1
Czech Republic	11	0.5
Denmark	16	0.7
Egypt	9	0.4
Europe	2	0.1
Finland	1	0.0
France	51	2.3
Germany	58	2.6
Ghana	1	0.0
Greece	21	0.9
Hong Kong	18	0.8
Hungary	15	0.7
India	33	1.5
Indonesia	11	0.5
Ireland	70	3.1
Italy	40	1.8
Japan	2	0.1
Jersey	2	0.1
Jordan	1	0.0
Kenya	6	0.3
Korea, Republic of	2	0.1
Latvia	2	0.1
Lebanon	2	0.1
Lithuania	3	0.1
Luxembourg	4	0.2
Macau	1	0.0
Malaysia	18	0.8
Malta	2	0.1
Mauritius	0	0.0
Mexico	2	0.1
Myanmar	2	0.1
Netherlands	32	1.4
New Zealand	7	0.3
Nigeria	12	0.5
Norway	5	0.2
Pakistan	2	0.1
Papua New Guinea	1	0.0
Peru	2	0.1
Philippines	9	0.4
Poland	12	0.5
Portugal	8	0.4

Romania	9	0.4
Russian Federation	15	0.7
Saudi Arabia	6	0.3
Serbia	1	0.0
Singapore	23	1.0
Slovakia	1	0.0
South Africa	21	0.9
Spain	49	2.2
Sri Lanka	1	0.0
Sweden	20	0.9
Switzerland	29	1.3
Taiwan	1	0.0
Thailand	5	0.2
Turkey	9	0.4
Ukraine	3	0.1
United Arab Emirates	11	0.5
United Kingdom	1291	57.5
United States	82	3.7
Vietnam	18	0.8
Prefer Not to Say	3	0.1

Table B1.1.1: Nationality profile of Norm 1

Ethnicity

Table B1.1.2 below shows the sample broken down by Ethnic Group. 34.0% of the sample indicated that they were White British.

Ethnic Group	Frequency (N)	Percent of Sample
WHITE		
White British	762	34.0
White Irish	123	5.5
Other White European	483	21.5
Other White background	102	4.5
MIXED		
White and Black Caribbean	21	0.9
White and Black African	29	1.3
White and Asian	37	1.6
Other Mixed background	25	1.1
ASIAN or ASIAN BRITISH		
Indian	166	7.4
Pakistani	36	1.6
Bangladeshi	16	0.7
Other Asian background	54	2.4

BLACK or BLACK BRITISH		
Black Caribbean	11	0.5
Black African	69	3.1
Other Black background	8	0.4
CHINESE or OTHER ETHNIC GROUP		
Chinese / East Asian	147	6.6
Other ethnic group	47	2.1
Prefer not to say	108	4.8

Table B1.1.2: Ethnicity profile of Norm 1

Employment Status

Table B1.1.3 below shows the sample broken down by Employment Status. 39.5% of the sample indicated that they were in Full-time Education.

Employment Status	Frequency (N)	Percent of Sample
Full time employment	611	27.2
Full time student	887	39.5
Other	52	2.3
Part time employment	134	6.0
Prefer not to say	67	3.0
Retired	14	0.6
Self-employed	62	2.8
Unemployed	417	18.6

Table B1.1.3: Employment Status profile of Norm 1

Industry Sector

Table B1.1.4 below shows the sample broken down by Industry Sector.

Industry Sector	Frequency (N)	Percent of Sample
Hospitality/Leisure	37	1.6
Advertising/Marketing	95	4.2
Automotive/Aerospace	66	2.9
Chemicals	51	2.3
Construction/Engineering	111	4.9
Education	79	3.5
Electronics	27	1.2
Entertainment/Media/Publishing	19	0.8
Finance - Investment Banking	149	6.6
Finance - Other	187	8.3
Finance - Retail Banking	34	1.5
FMCG	41	1.8
Government/Public Sector	42	1.9
Healthcare/Medical	43	1.9

HR Consultancy	34	1.5
Insurance	34	1.5
Legal	33	1.5
Logistics/Distribution	34	1.5
Manufacturing	46	2.0
Natural Resources	10	0.4
Non-Profit	10	0.4
Not Applicable	706	31.5
Other industry sector	51	2.3
Pharmaceutical	13	0.6
Prefer not to say	111	4.9
Professional Services	33	1.5
Property	8	0.4
Retail	48	2.1
Technology - Hardware	8	0.4
Technology - Services	16	0.7
Technology - Software	30	1.3
Telecommunications	14	0.6
Transportation/Travel	16	0.7
Utilities	7	0.3
Wholesale	1	0.0

Table B1.1.4: Industry Sector profile of Norm 1

Norm 2: Administrative, Operational, Apprentice and Non-graduate staff

Norm 2 comprises 629 Administrative, Operational, Apprentice and Non-graduate staff. The biographical data for this sample is shown below:

Gender

51.5% of the sample (n=324) was male and 33.1% (n=208) female. 15.4% of the sample (n=97) did not indicate his or her gender.

Nationality

Table B1.2.1 below shows the sample broken down by Country of Nationality. 53.1% of the sample indicated that they were UK Nationals.

Country of Nationality	Frequency (N)	Percent of Sample
Argentina	1	0.2
Australia	24	3.8
Bahrain	2	0.3
Belgium	5	0.8
Bolivia	2	0.3
Canada	8	1.3

China	2	0.3
Cyprus	4	0.6
Czech Republic	1	0.2
Denmark	9	1.4
Egypt	3	0.5
Europe	1	0.2
Finland	1	0.2
France	16	2.5
Germany	19	3.0
Greece	4	0.6
Hong Kong	7	1.1
Hungary	1	0.2
India	16	2.5
Indonesia	4	0.6
Ireland	14	2.2
Isle of Man	1	0.2
Italy	9	1.4
Jordan	1	0.2
Lithuania	3	0.5
Luxembourg	1	0.2
Malaysia	2	0.3
Mexico	1	0.2
Moldova, Republic of	1	0.2
Netherlands	11	1.7
New Zealand	4	0.6
Nigeria	4	0.6
Norway	2	0.3
Oman	2	0.3
Pakistan	1	0.2
Palestinian Territory	1	0.2
Philippines	5	0.8
Poland	6	1.0
Portugal	2	0.3
Romania	3	0.5
Russian Federation	7	1.1
Saudi Arabia	2	0.3
Singapore	7	1.1
South Africa	15	2.4
Spain	7	1.1
Sweden	8	1.3
Switzerland	5	0.8
Trinidad and Tobago	1	0.2
Turkey	1	0.2
Ukraine	4	0.6

United Kingdom	334	53.1
United States	23	3.7
Vietnam	9	1.4
Prefer Not to Say	2	0.3

Table B1.2.1: Nationality profile of Norm 2

Ethnicity

Table B1.2.2 below shows the sample broken down by Ethnic Group. 40.7% of the sample indicated that they were White British.

Ethnic Group	Frequency (N)	Percent of Sample
WHITE		
White British	256	40.7
White Irish	61	9.7
Other White European	69	11.0
Other White background	21	3.3
MIXED		
White and Black Caribbean	19	3.0
White and Black African	13	2.1
White and Asian	7	1.1
Other Mixed background	9	1.4
ASIAN or ASIAN BRITISH		
Indian	39	6.2
Pakistani	15	2.4
Bangladeshi	9	1.4
Other Asian background	16	2.5
BLACK or BLACK BRITISH		
Black Caribbean	17	2.7
Black African	6	1.0
Other Black background	3	0.5
CHINESE or OTHER ETHNIC GROUP		
Chinese / East Asian	21	3.3
Other ethnic group	37	5.9
Prefer not to say	11	1.7

Table B1.2.2: Ethnicity profile of Norm 2

Employment Status

Table B1.2.3 below shows the sample broken down by Employment Status. 32.9% of the sample indicated that they were in Full-time Employment.

Employment Status	Frequency (N)	Percent of Sample
Full time employment	207	32.9
Full time student	43	6.8
Other	39	6.2
Part time employment	109	17.3
Prefer not to say	33	5.2
Retired	9	1.4
Self-employed	58	9.2
Unemployed	131	20.8

Table B1.2.3: Employment Status profile of Norm 2

Industry Sector

Table B1.2.4 below shows the sample broken down by Industry Sector.

Industry Sector	Frequency (N)	Percent of Sample
Hospitality/Leisure	21	3.3
Advertising/Marketing	74	11.8
Automotive/Aerospace	60	9.5
Chemicals	29	4.6
Construction/Engineering	24	3.8
Education	34	5.4
Electronics	14	2.2
Entertainment/Media/Publishing	6	1.0
Finance - Investment Banking	10	1.6
Finance - Other	36	5.7
Finance - Retail Banking	15	2.4
FMCG	8	1.3
Government/Public Sector	12	1.9
Healthcare/Medical	23	3.7
HR Consultancy	14	2.2
Insurance	13	2.1
Legal	18	2.9
Logistics/Distribution	15	2.4
Manufacturing	14	2.2
Natural Resources	2	0.3
Non-Profit	8	1.3
Not Applicable	90	14.3
Other industry sector	5	0.8
Pharmaceutical	5	0.8
Prefer not to say	38	6.0
Professional Services	2	0.3
Property	1	0.2
Retail	22	3.5

Technology - Hardware	1	0.2
Technology - Software	1	0.2
Telecommunications	5	0.8
Transportation/Travel	3	0.5
Utilities	5	0.8
Wholesale	1	0.2

Table B1.2.4: Industry Sector profile of Norm 2

Insights Verbal

There are two norm groups currently available for Insights Verbal:

Norm	Description	Sample Size
1	Graduates, Professionals, Managers and Executives	3416
2	Administrative, Operational, Apprentice and Non-graduate staff	972

Information on each of these norm groups is presented in this section.

Norm 1: Graduates, Professionals, Managers and Executives

Norm 1 comprises 3416 Graduates, Professionals, Managers and Executives. The biographical data for this sample is shown below:

Gender

51.6% of the sample (n=1764) was male and 45.3% (n=1547) female. 3.1% of the sample (n=105) did not indicate his or her gender.

Nationality

Table B2.1.1 below shows the sample broken down by Country of Nationality. 52.1% of the sample indicated that they were UK Nationals.

Country of Nationality	Frequency (N)	Percent of Sample
Afghanistan	1	0.0
Albania	1	0.0
Argentina	2	0.1
Australia	153	4.5
Austria	5	0.1
Bangladesh	3	0.1
Belgium	15	0.4
Botswana	4	0.1
Brazil	2	0.1
Brunei Darussalam	8	0.2
Bulgaria	2	0.1
Cameroon	1	0.0
Canada	22	0.6
Cayman Islands	1	0.0
Chile	4	0.1
China	9	0.3

Colombia	2	0.1
Croatia	3	0.1
Cyprus	4	0.1
Denmark	20	0.6
Egypt	18	0.5
El Salvador	2	0.1
Estonia	3	0.1
Ethiopia	6	0.2
Europe	3	0.1
Finland	3	0.1
France	49	1.4
Georgia	1	0.0
Germany	29	0.8
Ghana	6	0.2
Greece	18	0.5
Guernsey	1	0.0
Haiti	3	0.1
Hong Kong	49	1.4
Hungary	7	0.2
India	173	5.1
Indonesia	17	0.5
Ireland	138	4.0
Italy	45	1.3
Japan	6	0.2
Kazakhstan	5	0.1
Kenya	13	0.4
Korea, Republic of	9	0.3
Kuwait	1	0.0
Latvia	1	0.0
Lebanon	4	0.1
Malaysia	10	0.3
Mali	1	0.0
Mauritius	1	0.0
Mexico	5	0.1
Morocco	2	0.1
Myanmar	1	0.0
Nepal	1	0.0
Netherlands	20	0.6
New Zealand	39	1.1
Nigeria	36	1.1
Norway	6	0.2
Oman	2	0.1
Pakistan	91	2.7
Panama	3	0.1

Philippines	70	2.0
Poland	4	0.1
Portugal	9	0.3
Qatar	4	0.1
Romania	16	0.5
Russian Federation	4	0.1
Saudi Arabia	21	0.6
Senegal	2	0.1
Serbia	4	0.1
Singapore	65	1.9
Slovenia	1	0.0
Somalia	2	0.1
South Africa	26	0.8
Spain	33	1.0
Sri Lanka	7	0.2
Sweden	16	0.5
Switzerland	13	0.4
Taiwan	2	0.1
Thailand	4	0.1
Trinidad and Tobago	1	0.0
Turkey	5	0.1
Uganda	2	0.1
Ukraine	1	0.0
United Arab Emirates	30	0.9
United Kingdom	1780	52.1
United States	164	4.8
Uzbekistan	1	0.0
Venezuela	2	0.1
Vietnam	29	0.8
Zambia	1	0.0
Zimbabwe	2	0.1
Prefer Not to Say	5	0.1

Table B2.1.1: Nationality profile of Norm 1

Ethnicity

Table B2.1.2 below shows the sample broken down by Ethnic Group. 35.1% of the sample indicated that they were White British.

Ethnic Group	Frequency (N)	Percent of Sample
WHITE		
White British	1199	35.1
White Irish	191	5.6

Other White European	426	12.5
Other White background	116	3.4
MIXED		
White and Black Caribbean	11	0.3
White and Black African	42	1.2
White and Asian	75	2.2
Other Mixed background	62	1.8
ASIAN or ASIAN BRITISH		
Indian	361	10.6
Pakistani	125	3.7
Bangladeshi	33	1.0
Other Asian background	122	3.6
BLACK or BLACK BRITISH		
Black Caribbean	28	0.8
Black African	145	4.2
Other Black background	8	0.2
CHINESE or OTHER ETHNIC GROUP		
Chinese / East Asian	227	6.6
Other ethnic group	62	1.8
Prefer not to say	183	5.4

Table B2.1.2: Ethnicity profile of Norm 1

Employment Status

Table B2.1.3 below shows the sample broken down by Employment Status. 37.5% of the sample indicated that they were in Full-time Employment.

Employment Status	Frequency (N)	Percent of Sample
Full time employment	1281	37.5
Full time student	915	26.8
Other	53	1.6
Part time employment	275	8.1
Prefer not to say	95	2.8
Retired	18	0.5
Self-employed	142	4.2
Unemployed	637	18.6

Table B2.1.3: Employment Status profile of Norm 1

Industry Sector

Table B2.1.4 shows the sample broken down by Industry Sector.

Industry Sector	Frequency (N)	Percent of Sample
Hospitality/Leisure	65	1.9
Advertising/Marketing	132	3.9
Automotive/Aerospace	72	2.1
Chemicals	46	1.3
Construction/Engineering	135	4.0
Education	218	6.4
Electronics	48	1.4
Entertainment/Media/Publishing	16	0.5
Finance - Investment Banking	136	4.0
Finance - Other	235	6.9
Finance - Retail Banking	47	1.4
FMCG	34	1.0
Government/Public Sector	183	5.4
Healthcare/Medical	112	3.3
HR Consultancy	59	1.7
Insurance	25	0.7
Legal	127	3.7
Logistics/Distribution	45	1.3
Manufacturing	59	1.7
Natural Resources	11	0.3
Non-Profit	35	1.0
Not Applicable	933	27.3
Other industry sector	104	3.0
Pharmaceutical	21	0.6
Prefer not to say	178	5.2
Professional Services	29	0.8
Property	11	0.3
Retail	81	2.4
Technology - Hardware	16	0.5
Technology - Services	33	1.0
Technology - Software	90	2.6
Telecommunications	35	1.0
Transportation/Travel	27	0.8
Utilities	14	0.4
Wholesale	4	0.1

Table B2.1.4: Industry Sector profile of Norm 1

Norm 2: Administrative, Operational, Apprentice and Non-graduate staff

Norm 2 comprises 972 Administrative, Operational, Apprentice and Non-graduate staff. The biographical data for this sample is shown below:

Gender

54.5% of the sample (n=530) was male and 38.3% (n=372) female. 7.2% of the sample (n=70) did not indicate his or her gender.

Nationality

Table B2.2.1 below shows the sample broken down by Country of Nationality. 57.5% of the sample indicated that they were UK Nationals.

Country of Nationality	Frequency (N)	Percent of Sample
Albania	1	0.1
Argentina	1	0.1
Australia	67	6.9
Austria	1	0.1
Azerbaijan	1	0.1
Belgium	4	0.4
Brazil	2	0.2
Canada	7	0.7
China	13	1.3
Croatia	1	0.1
Cyprus	1	0.1
Denmark	4	0.4
Egypt	12	1.2
Ethiopia	5	0.5
Europe	2	0.2
Faroe Islands	1	0.1
France	9	0.9
Germany	7	0.7
Ghana	1	0.1
Hong Kong	11	1.1
India	22	2.3
Indonesia	2	0.2
Ireland	39	4.0
Isle of Man	2	0.2
Italy	3	0.3
Kazakhstan	1	0.1
Korea, Republic of	3	0.3
Latvia	1	0.1
Lebanon	2	0.2
Lithuania	1	0.1

Macedonia	1	0.1
Malaysia	2	0.2
Malta	1	0.1
Mauritius	2	0.2
Morocco	1	0.1
Nepal	1	0.1
Netherlands	4	0.4
New Zealand	14	1.4
Nigeria	6	0.6
Norway	1	0.1
Pakistan	19	2.0
Philippines	26	2.7
Poland	1	0.1
Russian Federation	3	0.3
Saudi Arabia	10	1.0
Singapore	7	0.7
Slovakia	1	0.1
South Africa	24	2.5
Spain	7	0.7
Sri Lanka	1	0.1
Sweden	3	0.3
Switzerland	3	0.3
Turkey	2	0.2
United Arab Emirates	5	0.5
United Kingdom	559	57.5
United States	32	3.3
Uzbekistan	1	0.1
Vietnam	8	0.8

Table B2.2.1: Nationality profile of Norm 2

Ethnicity

Table B2.2.2 below shows the sample broken down by Ethnic Group. 44.7% of the sample indicated that they were White British.

Ethnic Group	Frequency (N)	Percent of Sample
WHITE		
White British	434	44.7
White Irish	68	7.0
Other White European	61	6.3
Other White background	23	2.4
MIXED		
White and Black Caribbean	22	2.3

White and Black African	20	2.1
White and Asian	25	2.6
Other Mixed background	20	2.1
ASIAN or ASIAN BRITISH		
Indian	56	5.8
Pakistani	35	3.6
Bangladeshi	15	1.5
Other Asian background	18	1.9
BLACK or BLACK BRITISH		
Black Caribbean	9	0.9
Black African	46	4.7
Other Black background	8	0.8
CHINESE or OTHER ETHNIC GROUP		
Chinese / East Asian	45	4.6
Other ethnic group	15	1.5
Prefer not to say	52	5.3

Table B2.2.2: Ethnicity profile of Norm 2

Employment Status

Table B2.2.3 below shows the sample broken down by Employment Status. 34.1% of the sample indicated that they were in Full-time Employment.

Employment Status	Frequency (N)	Percent of Sample
Full time employment	331	34.1
Full time student	136	14.0
Other	37	3.8
Part time employment	153	15.7
Prefer not to say	56	5.8
Retired	14	1.4
Self-employed	68	7.0
Unemployed	177	18.2

Table B2.2.3: Employment Status profile of Norm 2

Industry Sector

Table B2.2.4 below shows the sample broken down by Industry Sector.

Industry Sector	Frequency (N)	Percent of Sample
Hospitality/Leisure	36	3.7
Advertising/Marketing	66	6.8
Automotive/Aerospace	34	3.5
Chemicals	26	2.7
Construction/Engineering	42	4.3
Education	48	4.9
Electronics	23	2.4
Entertainment/Media/Publishing	21	2.2
Finance - Investment Banking	15	1.5
Finance - Other	57	5.9
Finance - Retail Banking	12	1.2
FMCG	7	0.7
Government/Public Sector	47	4.8
Healthcare/Medical	29	3.0
HR Consultancy	20	2.1
Insurance	12	1.2
Legal	24	2.5
Logistics/Distribution	21	2.2
Manufacturing	30	3.1
Natural Resources	2	0.2
Non-Profit	6	0.6
Not Applicable	198	20.4
Other industry sector	17	1.7
Pharmaceutical	2	0.2
Prefer not to say	64	6.6
Professional Services	9	0.9
Property	4	0.4
Retail	53	5.5
Technology - Hardware	4	0.4
Technology - Services	3	0.3
Technology - Software	11	1.1
Telecommunications	4	0.4
Transportation/Travel	20	2.1
Utilities	3	0.3
Wholesale	2	0.2

Table B2.2.4: Industry Sector profile of Norm 2

There are two norm groups currently available for Insights Inductive:

Norm	Description	Sample Size
1	Graduates, Professionals, Managers and Executives	3491
2	Administrative, Operational, Apprentice and Non-graduate staff	977

Information on each of these norm groups is presented in this section.

Norm 1: Graduates, Professionals, Managers and Executives

Norm 1 comprises 3491 Graduates, Professionals, Managers and Executives. The biographical data for this sample is shown below:

Gender

55.4% of the sample (n=1933) was male and 41.3% (n=1442) female. 3.3% of the sample (n=116) did not indicate his or her gender.

Nationality

Table B3.1.1 below shows the sample broken down by Country of Nationality. 30.1% of the sample indicated that they were UK Nationals.

Country of Nationality	Frequency (N)	Percent of Sample
Albania	17	0.5
Argentina	5	0.1
Asia/Pacific Region	4	0.1
Australia	318	9.1
Austria	7	0.2
Bahrain	1	0.0
Bangladesh	6	0.2
Belgium	24	0.7
Botswana	3	0.1
Brazil	22	0.6
Brunei Darussalam	30	0.9
Bulgaria	4	0.1
Cameroon	1	0.0
Canada	66	1.9
Cape Verde	3	0.1
Chile	15	0.4

China	21	0.6
Colombia	7	0.2
Croatia	3	0.1
Cyprus	6	0.2
Czech Republic	4	0.1
Denmark	27	0.8
Ecuador	2	0.1
Egypt	23	0.7
El Salvador	9	0.3
Estonia	3	0.1
Ethiopia	2	0.1
Europe	7	0.2
Fiji	1	0.0
Finland	7	0.2
France	69	2.0
Germany	39	1.1
Ghana	7	0.2
Greece	41	1.2
Guernsey	1	0.0
Hong Kong	34	1.0
Hungary	8	0.2
Iceland	6	0.2
India	249	7.1
Indonesia	25	0.7
Iran, Islamic Republic of	3	0.1
Ireland	28	0.8
Israel	3	0.1
Italy	42	1.2
Jordan	2	0.1
Kazakhstan	3	0.1
Kenya	7	0.2
Korea, Republic of	7	0.2
Kuwait	3	0.1
Latvia	5	0.1
Lebanon	6	0.2
Macedonia	9	0.3
Malaysia	43	1.2
Malta	3	0.1
Mauritius	1	0.0
Mexico	19	0.5
Mongolia	2	0.1
Morocco	3	0.1
Myanmar	1	0.0
Namibia	6	0.2

Nepal	1	0.0
Netherlands	132	3.8
New Zealand	41	1.2
Nigeria	34	1.0
Norway	38	1.1
Oman	3	0.1
Pakistan	13	0.4
Panama	1	0.0
Peru	2	0.1
Philippines	101	2.9
Poland	18	0.5
Portugal	18	0.5
Qatar	7	0.2
Romania	61	1.7
Russian Federation	14	0.4
Rwanda	1	0.0
Saudi Arabia	18	0.5
Serbia	2	0.1
Singapore	46	1.3
Slovakia	1	0.0
Slovenia	2	0.1
South Africa	64	1.8
Spain	35	1.0
Sri Lanka	1	0.0
Sweden	79	2.3
Switzerland	23	0.7
Taiwan	3	0.1
Thailand	26	0.7
Togo	2	0.1
Trinidad and Tobago	1	0.0
Tunisia	3	0.1
Turkey	35	1.0
Ukraine	3	0.1
United Arab Emirates	74	2.1
United Kingdom	1052	30.1
United States	188	5.4
Vietnam	12	0.3
Zimbabwe	11	0.3
Prefer Not to Say	2	0.1

Table B3.1.1: Nationality profile of Norm 1

Ethnicity

Table B3.1.2 below shows the sample broken down by Ethnic Group. 22% of the sample indicated that they were White British.

Ethnic Group	Frequency (N)	Percent of Sample
WHITE		
White British	770	22
White Irish	82	2
Other White European	823	23
Other White background	173	4.9
MIXED		
White and Black Caribbean	19	0.5
White and Black African	37	1.1
White and Asian	88	2.5
Other Mixed background	55	1.6
ASIAN or ASIAN BRITISH		
Indian	484	13.9
Pakistani	83	2.4
Bangladeshi	32	0.9
Other Asian background	125	3.6
BLACK or BLACK BRITISH		
Black Caribbean	10	0.3
Black African	123	3.5
Other Black background	11	0.3
CHINESE or OTHER ETHNIC GROUP		
Chinese / East Asian	303	8.7
Other ethnic group	87	2.5
Prefer not to say	186	5.3

Table B3.1.2: Ethnicity profile of Norm 1

Employment Status

Table B3.1.3 below shows the sample broken down by Employment Status. 41.9% of the sample indicated that they were in Full-time Employment.

Employment Status	Frequency (N)	Percent of Sample
Full time employment	1462	41.9
Full time student	797	22.8
Other	87	2.5
Part time employment	194	5.6
Prefer not to say	134	3.8
Retired	15	0.4

Self-employed	141	4.0
Unemployed	661	18.9

Table B3.1.3: Employment Status profile of Norm 1

Industry Sector

Table B3.1.4 below shows the sample broken down by Industry Sector.

Industry Sector	Frequency (N)	Percent of Sample
Hospitality/Leisure	55	1.6
Advertising/Marketing	181	5.2
Automotive/Aerospace	122	3.5
Chemicals	56	1.6
Construction/Engineering	171	4.9
Education	167	4.8
Electronics	71	2.0
Entertainment/Media/Publishing	27	0.8
Finance - Investment Banking	185	5.3
Finance - Other	270	7.7
Finance - Retail Banking	75	2.1
FMCG	45	1.3
Government/Public Sector	119	3.4
Healthcare/Medical	95	2.7
HR Consultancy	56	1.6
Insurance	42	1.2
Legal	101	2.9
Logistics/Distribution	70	2.0
Manufacturing	65	1.9
Natural Resources	7	0.2
Non-Profit	19	0.5
Not Applicable	735	21.1
Other industry sector	67	1.9
Pharmaceutical	28	0.8
Prefer not to say	226	6.5
Professional Services	39	1.1
Property	23	0.7
Retail	56	1.6
Technology - Hardware	13	0.4
Technology - Services	39	1.1
Technology - Software	139	4.0
Telecommunications	64	1.8
Transportation/Travel	45	1.3
Utilities	10	0.3
Wholesale	8	0.2

Table B3.1.4: Industry Sector profile of Norm 1

Norm 2: Administrative, Operational, Apprentice and Non-graduate staff

Norm 2 comprises 977 Administrative, Operational, Apprentice and Non-graduate staff. The biographical data for this sample is shown below:

Gender

54.8% of the sample (n=535) was male and 35.2% (n=344) female. 10% of the sample (n=98) did not indicate his or her gender.

Nationality

Table B3.2.1 below shows the sample broken down by Country of Nationality. 31.1% of the sample indicated that they were UK Nationals.

Country of Nationality	Frequency (N)	Percent of Sample
Algeria	1	0.1
Australia	95	9.7
Austria	5	0.5
Bangladesh	1	0.1
Belgium	4	0.4
Bolivia	1	0.1
Botswana	1	0.1
Brazil	3	0.3
Brunei Darussalam	1	0.1
Bulgaria	13	1.3
Cameroon	7	0.7
Canada	32	3.3
Chile	4	0.4
China	2	0.2
Colombia	3	0.3
Costa Rica	3	0.3
Croatia	1	0.1
Cyprus	1	0.1
Czech Republic	4	0.4
Denmark	8	0.8
Ecuador	1	0.1
Egypt	9	0.9
El Salvador	1	0.1
Europe	2	0.2
Finland	4	0.4
France	18	1.8
Germany	14	1.4
Greece	7	0.7
Guernsey	1	0.1
Hong Kong	13	1.3

India	26	2.7
Indonesia	3	0.3
Ireland	10	1.0
Israel	1	0.1
Italy	6	0.6
Kazakhstan	1	0.1
Kenya	3	0.3
Latvia	3	0.3
Lebanon	4	0.4
Malaysia	10	1.0
Malta	2	0.2
Mexico	6	0.6
Mozambique	3	0.3
Nepal	1	0.1
Netherlands	21	2.1
New Zealand	18	1.8
Nigeria	8	0.8
Norway	9	0.9
Oman	3	0.3
Pakistan	4	0.4
Philippines	44	4.5
Poland	7	0.7
Qatar	2	0.2
Romania	10	1.0
Russian Federation	4	0.4
Rwanda	3	0.3
Saudi Arabia	9	0.9
Singapore	10	1.0
South Africa	24	2.5
Spain	3	0.3
Sweden	35	3.6
Switzerland	8	0.8
Taiwan	7	0.7
Thailand	6	0.6
Turkey	10	1.0
United Arab Emirates	26	2.7
United Kingdom	304	31.1
United States	62	6.3
Vietnam	1	0.1

Table B3.2.1: Nationality profile of Norm 2

Ethnicity

Table B3.2.2 below shows the sample broken down by Ethnic Group. 36.1% of the sample indicated that they were White British.

Ethnic Group	Frequency (N)	Percent of Sample
WHITE		
White British	353	36.1
White Irish	90	9.2
Other White European	39	4.0
Other White background	129	13.2
MIXED		
White and Black Caribbean	11	1.1
White and Black African	18	1.8
White and Asian	14	1.4
Other Mixed background	14	1.4
ASIAN or ASIAN BRITISH		
Indian	65	6.7
Pakistani	14	1.4
Bangladeshi	11	1.1
Other Asian background	35	3.6
BLACK or BLACK BRITISH		
Black Caribbean	17	1.7
Black African	41	4.2
Other Black background	7	0.7
CHINESE or OTHER ETHNIC GROUP		
Chinese / East Asian	43	4.4
Other ethnic group	25	2.6
Prefer not to say	51	5.2

Table B3.2.2: Ethnicity profile of Norm 2

Employment Status

Table B3.2.3 below shows the sample broken down by Employment Status. 34.8% of the sample indicated that they were in Full-time Employment.

Employment Status	Frequency (N)	Percent of Sample
Full time employment	340	34.8
Full time student	75	7.7
Other	52	5.3
Part time employment	173	17.7

Prefer not to say	46	4.7
Retired	23	2.4
Self-employed	106	10.8
Unemployed	162	16.6

Table B3.2.3: Employment Status profile of Norm 2

Industry Sector

Table B3.2.4 below shows the sample broken down by Industry Sector.

Industry Sector	Frequency (N)	Percent of Sample
Hospitality/Leisure	38	3.9
Advertising/Marketing	122	12.5
Automotive/Aerospace	71	7.3
Chemicals	52	5.3
Construction/Engineering	56	5.7
Education	47	4.8
Electronics	22	2.3
Entertainment/Media/Publishing	18	1.8
Finance - Investment Banking	24	2.5
Finance - Other	32	3.3
Finance - Retail Banking	14	1.4
FMCG	11	1.1
Government/Public Sector	24	2.5
Healthcare/Medical	31	3.2
HR Consultancy	14	1.4
Insurance	8	0.8
Legal	39	4.0
Logistics/Distribution	26	2.7
Manufacturing	28	2.9
Natural Resources	1	0.1
Non-Profit	2	0.2
Not Applicable	145	14.8
Other industry sector	24	2.5
Pharmaceutical	3	0.3
Prefer not to say	41	4.2
Professional Services	10	1.0
Property	6	0.6
Retail	25	2.6
Technology - Hardware	6	0.6
Technology - Services	2	0.2
Technology - Software	6	0.6
Telecommunications	1	0.1
Transportation/Travel	19	1.9

Utilities	2	0.2
Wholesale	7	0.7

Table B3.2.4: Industry Sector profile of Norm 2

Appendix C: Insights Series Development Timeline

Insights Series Development Timeline															
Tasks	**Week 1**	**Week 2**	**Week 3**	**Week 4**	**Week 5**	**Week 6**	**Week 7**	**Week 8**	**Week 9**	**Week 10**	**Week 11**	**Week 12**	**Week 13**	**Week 14**	**Ongoing**
Stage 1: Scoping															
Define & agree scope															
Job Analysis															
Stage 2: Write Test Items															
Write Example Test Items															
Review & Amend Test Item Templates															
Sign Off On Amended Item Templates															
Commence Full Scale Item Writing															
Stage 3: Item Calibration															
Review Written Items before Trialling															
Trialling & Data Collection for Calibration, Reliability, Adverse Impact															
Item Amendment & Re-Trialling															
Ongoing Review & Analysis of Items in Trial															
Item Amendment & Re-Trialling															
Stage 5: Norm Group Construction															
Define Norms															
Collect Norm Group Data															
Review Norm Definitions															
Construct Norms															
Stage 6: Analysis															
Analyse Final Data															
Finalise & Create Item Banks															
Adverse Impact Analyses															
Stage 7: Validation															
Validation Research															
Analyse Validity Data															
Rollout of Selection Tools															
Ongoing Item Bank Development															