

TECHNICAL MANUAL

&

USER GUIDE

SMART

This manual has been created by the psychologists from the Science and Innovation department of Central Test.

The assessment **SMART** and the Central Test logo are registered trademarks.

No reproduction, modification or adaptation is permitted without the written consent of the publisher Central Test.

Table of Contents

1. Introduction	3
2. General Description and Applicability	5
2.1. Main Areas of Application	5
2.2. Questionnaire Structure	6
2.3. Mode of Administration, Languages and Special Requirements	6
2.4. Test Security	7
2.5. Administration Instructions	7
2.6. Technical and Security Issues	8
2.6.1. Respecting the Ethical Charter	8
2.6.2. Confidentiality and Security of User Data	9
2.6.3. Hardware Architecture and Server Availability	9
2.6.4. Computer-generated Reports	10
2.7. Supply, Conditions and Availability	10
3. Conceptual Framework and Development	11
3.1. History of Intelligence (GMA) and Aptitude Tests	11
3.2. Aptitude Tests at Work	13
3.3. History of Computerised Adaptive Testing (CAT)	13
3.4. Calibrated Item Bank	13
3.5. Item Response Theory	14
3.5.1. Item Parameters	14
3.5.1.a. Item Difficulty and Item Discrimination	15
3.5.2. Advantages and Validation Studies	16
4. SMART Modules	18
4.1. Logical Reasoning	18
4.1.1. Development	18
4.1.2. Norms	20
4.1.3. Reliability	21
4.1.3.a. Internal Reliability	22

4.1.4. Item Parameters	22
4.2. Numerical Reasoning	24
4.2.1. Development	24
4.2.2. Norms	25
4.2.3. Reliability	27
4.2.3.a. Internal Reliability	27
4.2.4. Item Parameters	28
4.3. Verbal Reasoning	30
4.3.1. Development	30
4.3.2. Norms	31
4.3.3. Reliability	33
4.3.3.a. Internal Reliability	33
4.3.4. Item Parameters	34
5. Scoring	36
5.1. Percentile Rank Scores and the Scoring Method	36
5.2. Combined Score	36
References	38

1. Introduction

What exactly is intelligence? Researchers have defined the concept of intelligence in many different ways throughout the history of psychology. English psychologist Charles Spearman believed that intelligence consists of one underlying factor of general mental ability, conceptualised as the g factor, which can be measured and compared among individuals. In his Two-Factor Theory of Intelligence (1927), Spearman theorised that all factors of intelligence consist of specific “s” factors that can be summarised by the general g factor of intelligence. These specific “s” factors refer to separate, single and specific cognitive abilities.

In the 1940s, Raymond Cattell proposed a theory of intelligence that divides general intelligence into two components: crystallised intelligence and fluid intelligence. Crystallised intelligence is characterised as acquired knowledge and the ability to retrieve it. When an individual learns, remembers, and recalls information, he / she is using crystallised intelligence. Fluid intelligence encompasses the ability to see complex relationships and solve problems. Fluid intelligence helps an individual tackle complex, abstract challenges in their daily life, whereas crystallised intelligence helps to overcome concrete, straightforward problems (Cattell, 1963).

SMART is an adaptive assessment tool designed to measure an individual’s specific reasoning skills (as theorised by Spearman’s “s” factors). The tool also combines the results from the specific reasoning aptitudes to provide an overall score on the reasoning abilities (related to Spearman’s “g” factor). **SMART** measures Crystallised Intelligence via the **SMART - VERBAL & SMART - NUMERICAL** Modules, and measures Fluid Intelligence via the **SMART - LOGICAL** Module.

Unlike traditional reasoning tests which typically involve test-takers being given the same set of questions - some harder and some easier - and then comparing performance between individuals, an assessment like **SMART** provides a more accurate representation of the individual’s skills due to its adaptive nature. This means that **SMART**’s measurement parameters will adjust according to the user’s response patterns throughout the test.

As an adaptive test, **SMART** responds to the test-takers ability, and assesses each individual at the limits of their capabilities. When individuals get questions right, they’re presented with increasingly more difficult questions. When individuals get questions wrong, they’re presented with less difficult questions. In this way, an adaptive test like **SMART** allows us to hone in on the ability-level of the person being tested.

The development and administration of an adaptive test is decidedly more challenging than a traditional fixed form test; it requires a great deal more technical coordination and expertise

in psychometric techniques and analysis. However, adaptive tests like **SMART** provide distinct advantages that make them entirely worth the effort.

The fact that **SMART** adapts to the test-taker means that their experience is unique and close to their ability level. They will rarely be given items that are frustratingly difficult or waste time on items that are too easy, meaning that test-takers will have a more positive experience of the assessment process.

This tailored experience also means that the integrity and security of the assessment questions can be maintained for longer, compared to fixed-form tests where every participant sees every question. What's more is that we can continue to add to the pool of items, thereby improving the accuracy and the security of the overall test.

SMART is an updated, adaptive version of Central Test's *Reasoning Test*.

The following are the leverages a user of **SMART** benefits from:

- Test items are accurate, representative, consistent, clear, fair and culturally-free.
- Test items are created in various levels of difficulty and as the candidate begins to respond, the upcoming items are calibrated to fit the response pattern of the candidate.
- Candidates are less likely to be frustrated by excessively difficult items, or waste time on items far below their ability level.
- A score is obtained based on the level of 'difficulty', which the test decides a candidate is best suited to.
- Developed using a state-of-the-art psychometric technique, the Item Response Theory.
- User-friendly and customisable reports that can be easily tailored by the administrator.

In this manual and user guide you will find useful information about the questionnaire development, theoretical background, psychometric properties as well as practical applications of **SMART**.

2. General Description and Applicability

Wechsler (1944) describes intelligence as “the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment.” **SMART** assesses the overall and specific intellectual capacities of an individual across a range of domain specific areas. Furthermore, the technical foundations of **SMART** allow for the addition of additional reasoning tasks where required.

2.1. Main Areas of Application

SMART was specifically designed to help HR professionals, managers and business practitioners recruit, select and develop employees across the globe. It may be used throughout an individual's career life-span and is suitable for all levels. Table 1 below details the different applications of **SMART**.

Table 1. Applications of SMART.

Recruitment	Internal Mobility	Selection in Institutions
To recruit individuals with high intellectual abilities, which in turn leads to better returns on investment for your organisation.	To identify your employees' reasoning abilities, thereby allowing you to predict which areas they will be most likely to excel in, and thus validate their potential to succeed on tasks and in future projects.	To give a clear picture of a student's aptitude and learning ability, which in turn helps in selecting candidates at your college or B-school.

SMART can also be used in conjunction with other occupational-based assessments to acquire a more comprehensive profile of individuals in specific roles. For example, combining it with the **Professional Profile 2©** provides a more precise and exhaustive profile of the individual, as displayed in Table 2.

Table 2. Application Example of **SMART** with Another Assessment.

Additional Assessment	Objective	Advantage
Professional Profile 2©	Career Development	In-depth information about intellectual abilities and personality fit, tailored for a specific career

2.2. Questionnaire Structure

SMART, a computer-adaptive assessment, consists of items selected from a collection of items, known as an item bank. The items are chosen to match the estimated level of reasoning aptitude of the test-taker. If the test-taker succeeds on an item, a slightly more challenging item is presented next, and vice-versa; thereby, quickly converging into a sequence of items bracketing and converging on the test-taker's effective aptitude level.

2.3. Mode of Administration, Languages and Special Requirements

SMART is administered online and works with all modern computer operating systems including Windows, MAC-OS and Linux. Like other modern questionnaires, the assessment can be administered remotely and without supervision. However, should standardised group assessments be required, this can easily be accommodated. Reports are solely generated online.

The questionnaire is currently available in English, French, and Spanish.

SMART is designed to be easily used by individuals from diverse backgrounds and by those with a range of special needs. All test-takers are able to adjust visual screen specifications based on individual requirements. For example, respondents with visual impairments or those requiring a larger font size for comfort, may increase the font size by enlarging the computer screen's display (by pressing "CTRL" and "+"). Similarly, respondents are able to access test instructions throughout the course of the assessment.

Administrators are encouraged to contact the test publisher for guidance on administering the questionnaire to individuals or groups with special requirements.

2.4. Test Security

An adaptive test like **SMART** is highly resistant to being compromised. The likelihood that two people will take the same test is very low, thus pre-learning answers or copying from another test taker is not possible.

Test security is improved by randomizing the item selection and using performance on these items to estimate ability. This also helps to ensure that a wide range of test items are used, further enhancing the security of the test. Randomisation also prevents test-takers from experiencing similar items. Additionally, randomization also equalizes the usage of the item bank.

Test security is also improved because item exposure is limited. Any item can only be seen by a fraction of the candidates, at most. Thus, the chances of a large number of test-takers experiencing identical tests is diminished overall. Furthermore, due to overlap, item difficulties can be compared and monitored, thus ensuring a fair evaluation of performance for all test-takers (Chae et al., 2000).

2.5. Administration Instructions

In most cases, the assessment will be taken remotely through online access and without supervision. However, in the case of taking the assessment within a supervised setting, the following guidelines should be applied.

Although instructions for completing **SMART** are presented on-screen, it is important that administrators communicate with the candidates. Prior to starting the assessment, administrators should verify that candidates understand what they are required to do and that candidates are given the opportunity to seek clarification regarding the assessment.

Test instructions are the same for every individual. Standardised instructions ensure that all participants are exposed to similar conditions during the assessment process, to ensure that it is fair and objective, and that assessment conditions do not directly or indirectly give any individual or group an advantage.

Further tips:

- If candidates have questions after the assessment session has begun, the administrator should try to answer them without disturbing other candidates.

- If a candidate does not understand a question, the administrator should answer carefully and avoid explaining the meaning of words or questions in his/her own words in order to not influence the candidate.
- If an administrator suspects that a candidate's lack of understanding is due to language proficiency, the administrator should ask the candidate to take the assessment in the language they are most comfortable with.

Candidates receive a notification to take the questionnaire via email with an embedded web link and login details. Web links to the questionnaire are password protected and unique to the candidate's account. It is possible to set expiration dates in order to ensure that candidates complete the questionnaire within the specified time-frame. It is possible to configure the invitation in the way that the assessment must be completed at the same IP address once the questionnaire has been accessed from a particular IP address. Completion times are also recorded for monitoring purposes in order to allow administrators to see if a candidate rushed through the questionnaire.

2.6. Technical and Security Issues

This chapter details guidelines governing the security of administering **SMART**. These guidelines ensure that every administration of **SMART** is a pleasant, objective and fair process. Security specifications ensure that candidate information is protected both during administration and after assessment completion.

2.6.1. Respecting the Ethical Charter

As a publisher of online assessment tools, Central Test respects the legal conventions of French Labour Law and the regulatory provisions resulting from the law of 6 January 1978 pertaining to information technology, files and freedoms: Declaration N°1834067.

Additionally, as of 25th May, 2018, Central Test complies with the GDPR (General Data Protection Regulation) regulations made by the European Parliament & Council.

Test results and personal data cannot be shared with third parties without explicit consent from the candidate. Candidates have the right to access their results and the assessor must grant them access when requested. Furthermore, if the candidate requests it, all data must be permanently deleted; to do so the candidate must contact either the organisation or Central Test directly.

2.6.2. Confidentiality and Security of User Data

Central Test guarantees the security and confidentiality of data generated during the administration of online tests. Access codes are encrypted in the database and cannot be accessed by anyone. Passwords are computationally generated and are not visible to companies. A URL rewriting module (secure link generator) is installed and has transparent parameter passing for the user.

Browsing and data exchange via the Central Test application can be secured and encrypted using SSL. Secure browsing is guaranteed by the https protocol, which is represented by an icon in the form of a padlock on the internet browser. The Central Test application has a "GeoTrust SSL True BusinessID" digital signature, issued and certified by GeoTrust©. This certificate is irrefutable evidence of our serious attitude towards security and protects all Central Test assessments from phishing attacks. Invitations to take tests are made via secure links that are of limited use (in terms of numbers and duration). The tests are timed and IP addresses are tracked. Access to questionnaires may be limited in terms of numbers and duration.

2.6.3. Hardware Architecture and Server Availability

Central Test's architecture consists of a network of 10 physical dedicated servers providing high-quality service.

Each server is housed in a specialised data centre featuring:

- Dedicated lines
- A guaranteed bandwidth of 100Mbps (per server)
- Redundant energy sources
- 4-hour Equipment Restoration time
- 2-hour Incident Intervention time
- Monitoring and assistance 24/7/365

Access to the server administration is restricted to an IP address and is password-secured. Data exchanges are secure.

Central Test's application has been tested on and is compatible with the following browsers:

- Internet Explorer 8 or above
- Firefox 21 or above
- Google Chrome 30 or above

- Safari 7 or above

2.6.4. Computer-generated Reports

Reports can be generated within minutes of candidates completing the questionnaire. The **SMART** report is user friendly and customisable. Report content can be tailored depending on the needs of the administrator. Separate reports may be configured for the candidate and the company, recruiter, coach or psychologist. Reports are available in PDF format.

2.7. Supply, Conditions and Availability

Conditions of supply and availability for **SMART** vary from country to country. However, the following points generally apply:

- **SMART** is only available for Psychologists, HR practitioners and coaches to administer.
- Training is compulsory prior to use and distribution.
- This manual is available to individuals who have been trained to use **SMART**.
- **SMART** can be ordered online, or by contacting any Central Test customer service representative or local partner worldwide (customer@centraltest.com).
- For information on **SMART** prices and related products and services, contact any Central Test customer service representative or local partner worldwide.
- For more information on the training, please contact Central Test's customer service or one of Central Test's partners.

3. Conceptual Framework and Development

3.1. History of Intelligence (GMA) and Aptitude Tests

Human intelligence allows us to learn from our surroundings and experiences, solve puzzles, comprehend abstract notions, and understand the information that we are given. Providing a reliable measure of intelligence is one of the most researched areas of Psychology, and one of its first origins is said to be the work of Galton in 1869, where he used an array of primitive tests to measure the motor skills of a large sample of people. His work led to the development of aptitude testing as it is now in psychometrics.

Although there is still debate as to the exact definition of mental ability, there is a large consensus among researchers that intelligence can be defined as the ability to process and understand *new* information; and aptitude tests such as **SMART** allow for an objective measurement of an individual's aptitude on aspects such as verbal, numerical and logical reasoning abilities, thereby enabling a method of assessing the individual's intelligence.

Charles Spearman's Two-Factor Theory of Intelligence

Psychologist Charles Spearman defined general intelligence as "the ability to infer relationships and, from them, to make correlations".

He postulated the Two-Factor Theory of intelligence in 1927 as a theory of trait organisation based on the statistical analysis of IQ test scores. Spearman advanced the theory that all intellectual factors have a single factor which is called the *g* factor¹ as well as a number of specific or "*s*" factors. These specific factors refer to separate, single and specific activities and abilities. He explored the statistical interrelations of scores on various tests and concluded that a positive correlation between any two mental functions must be attributed to the *g* factor, since "*s*" factors have low correlations among them. Spearman reasoned that the *g* factor runs through all mental abilities and predicts an individual's performance; while each measure of the specific ("*s*") factors operates in only a single activity. Spearman proposed that the *g* factor is a sort of "mirror" or reflection of one's intelligence²; and he hypothesised that *g*, i.e., "general intelligence", was the primary reason for positive relations among the various scores on IQ tests.

Charles Spearman devised this Two-Factor Theory to define intelligence, using a statistical technique known as "factor analysis". This technique is derived from the findings that an individual who does well on one mental ability test, tended to also perform well on others, while those that performed poorly on one test also had a tendency to not do so well on others. As a result, Spearman's theory looked at patterns in an individual's scores throughout an

assessment, and proposed that there are two underlying factors that can explain individual differences within the scores; the “general” factor (general intelligence) and the “specific” factor (intelligence related specifically to the test in question). This idea is now commonly referred to as General Mental Ability (GMA), and is a construct that is widely used to this day, with measurements such as IQ scores.

In a group of IQ tests, the test that measures most accurately g is the test that has the highest correlations with all the others. g loading is the correlation between a test and g in which the score reflects general mental ability. Most g loaded tests have at least one form of abstract reasoning. A common, but still unproven view is that, perhaps, g is the genetically-determined essence of intelligence (Hally, 2012).

Raymond Cattell's Theory of Fluid and Crystallised Intelligence

In the 1940s, Raymond Cattell proposed a theory of intelligence that divided general intelligence into two components: crystallised intelligence and fluid intelligence. Crystallised intelligence is characterised as acquired knowledge and the ability to retrieve it. When an individual learns, remembers, and recalls information, he / she is using crystallised intelligence. Fluid intelligence encompasses the ability to see complex relationships and solve problems. Fluid intelligence helps an individual tackle complex, abstract challenges in their daily life, whereas crystallised intelligence helps to overcome concrete, straightforward problems (Cattell, 1963).

In his 1987 book, “Intelligence: Its Structure, Growth, and Action”, Cattell referred to the ability to reason as fluid intelligence because it “has the ‘fluid’ quality of being directable to almost any problem” (Vinney, 2020). This ability is thought to be unaffected by environmental factors such as quality of education, and differs from what Cattell labelled crystallised intelligence; referring to previously learnt knowledge, which, in contrast, is affected by environmental factors. He referred to knowledge acquisition as crystallised intelligence because it “is invested in the particular areas of crystallised skills which can be upset individually without affecting the others” (Vinney, 2020).

Fluid intelligence is largely measured through tests measuring logical and abstract reasoning skills. Crystallised intelligence on the other hand, is often measured using verbal and numerical aptitude tests. Although fluid and crystallised intelligence represent two distinct sets of abilities, they can and often do work together. Fluid intelligence is often used when learning new things. When you encounter a new subject, you use your fluid intelligence to understand the material through logic and analysis. Once you understand the material, the information will be incorporated into your long-term memory, where it can develop into crystallised knowledge. For example, when cooking a meal, one uses crystallised intelligence to understand and follow the instructions in a recipe, and uses fluid intelligence when modifying

spices and other ingredients to suit their tastes or dietary requirements. Similarly, when taking a math test, the formulas and math knowledge (like the meaning of a plus sign) comes from crystallised intelligence. The ability to develop a strategy to complete a complicated problem, on the other hand, is the product of fluid intelligence (Vinney, 2020).

3.2. Aptitude Tests at Work

Within the recruitment process, aptitude assessments are vital. They allow the recruiter to objectively decipher whether a candidate is capable of understanding and performing a skill set specific to the role. Due to their objective nature, aptitude tests also provide a comparison between candidates, by looking at their scores.

3.3. History of Computerised Adaptive Testing (CAT)

Computerised Adaptive Testing (CAT) was developed largely in the 1970's and 1980's, as it became accessible to process large data sets on computers, and psychometrically analyse such data (Weiss, 1983). The main purpose of CAT is to take into account the wide range of ability levels of the assessment-taking audience, and therefore adapt to these levels by selecting from an item bank, test items that are appropriate to the user's trait level (Meijer, 1999).

Computerised Adaptive Screening Test (Gade, Johnson & Pliske, 1984) was later implemented for use by the army, with the aim of accurately judging an applicant's ability estimate, and therefore helping to determine whether applicants needed further testing. MicroCAT was also introduced in 1984 by the Assessment Systems Corporation, providing users with a way of self-development, and administering computerised test banks (Rocklin, Roos, Wise, & Yoes, 1996).

Providing advantages such as reduced user fatigue, costs, time and more accurate data, CAT continues to be used widely to the present day.

Thompson and Weiss (2011) lay out the five components that are proposed to make up CAT. They are; 1. Calibrated item bank, 2. Starting point, 3. Item selection algorithm, 4. Scoring algorithm, and 5. Termination criterion.

3.4. Calibrated Item Bank

The bank of items from which test questions are chosen is referred to as the Calibrated Item Bank. This means that the defined parameters of each item are stored on the computer.

Important points to take into account when creating the item pool, are that the bank should be as large as possible (Flaughner, 2000), in order to ensure the reliability of the parameters generated.

Item parameters were determined through a pretesting and test equating procedure, whereby smaller lists that contained both 'unique' and 'common' items were administered to a large sample of participants. Item difficulty parameters were calculated for each of the lists. After the parameters were calculated, anchor items were used to derive equating constants for each list via mean / sigma equating procedures (Kolen & Brennan, 1995).

3.5. Item Response Theory

Item Response Theory (IRT) was first proposed in the field of psychometrics for the scoring of assessments measuring ability, and over the past few decades has grown significantly and is now widely accepted as an improvement over Classical Test Theory (CTT) for the assessment of both aptitude and personality (Crocker & Algina, 1986). The theory represents a broad family of statistical models providing an explanation for the relationship between latent traits (unobservable characteristics), and observable outcomes. In psychometric testing, models of IRT aim to establish a link between individual items in an assessment, the individuals responding to the test items, and the underlying trait being measured (e.g., numerical reasoning ability).

IRT for the assessment of ability relies upon four assumptions. The first assumption, *monotonicity*, indicates that as a trait level is increasing, the probability of a correct response also increases. The second, *unidimensionality*, assumes that there is one dominant latent trait being measured, and it is this trait which is directing the responses observed for each item in the assessment. Thirdly, *local independence* assumes that the responses given to individual items in a test are collectively independent with a certain level of ability. Finally, the assumption of *invariance* allows us to estimate the item parameters from any position on the item response curve. Thus, we can estimate the parameters of each item from any group of test-takers who have responded to the item. These four assumptions, if maintained, allow us to conclude that observed differences in correct responses between test-takers will be due to differences in their level of a latent trait (e.g., numerical reasoning ability).

3.5.1. Item Parameters

Item parameters refer to an item's characteristics and position on the latent trait continuum, allowing IRT models to predict test-takers' answers, with the assumption that every response to an item on a measure provides an indication of an individual's level of the latent trait/ability (or θ). This level of ability, or θ , can therefore be summarised as the individual's probability of

selecting the correct answer for a given item. The higher the θ , the more likely an individual is to select the correct answer. In other words, at each ability level, there will be a certain probability that an individual with that ability will give a correct answer to the item.

In the 2-parameter logistic (2PL) model of IRT this probability is calculated by the given formula:

$$P(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-a(\theta-b)}}$$

Where:

b is the difficulty parameter

a is the discrimination parameter

$L = a(\theta - b)$ is the logistic deviate (logit)

θ is an ability level.

This relationship can be depicted graphically using an Item Characteristic Curve (ICC). In statistical applications, θ ranges between -3 and 3 (see Figure 1).

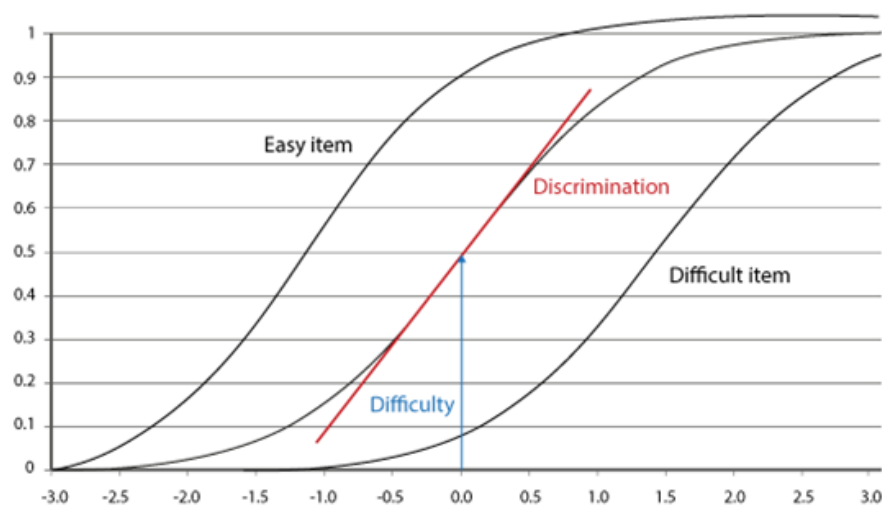


Figure 1. Item Characteristic Curves of an Easy, Moderate and Difficult Item from an Ability Test.

3.5.1.a. Item Difficulty and Item Discrimination

Under the 2PL model of IRT, item characteristic curves contain two properties: *Item difficulty*, and *item discrimination*. *Item difficulty* indicates where the item functions along the ability scale, calculated at the point of median probability (the ability at which 50% of test-takers select the correct answer). On an ICC, as shown above, difficult items shift towards the right

of the scale, indicating the higher level of ability of individuals who answered the item correctly. Easier items shift towards the left of the scale.

Item discrimination indicates the rate at which the probability of selecting the correct response changes at a given ability level (Baker, 2001). On an ICC, the item discrimination is the slope of each curve: the steeper the slope, the greater the discrimination of the item (see Figure 1). This is particularly important when comparing individuals who exhibit similar levels of the latent trait - that is, items with higher discrimination are needed in order to more accurately place individuals along the continuum of the latent trait being measured.

3.5.2. Advantages and Validation Studies

IRT holds many benefits for measurement accuracy and reliability in psychometric testing. The main feature differentiating IRT from previous theory is that reliability and error measurement are computed and can therefore be assessed for each item rather than for the test as a whole, allowing for greater accuracy of item selection in test construction with fewer test items required (Thomas, 2011). This is known as *item information function* at the item level, and *test information function* at the test level.

A key advantage of IRT is that it allows for computerised adaptive testing, as is used in **SMART**. IRT adaptive tests function by making an initial estimate of the test-taker's ability, and then repeatedly selecting and displaying items from the item bank which will provide maximal information about the test-taker's true ability. The test calculates a new estimate of the individual's ability based on their responses, and terminates the test when it reaches termination criteria (e.g., test time limit or question limit). This enables greater test efficiency as maximal information regarding an individual's ability can be obtained in a shorter test duration, reducing testing time and burden whilst improving reliability (Vispoel et al., 1999; Thomas, 2011).

Another advantage of IRT is the assumption of invariance of item parameters; item parameters do not depend on the individual's ability, and the individual's parameters do not depend on the items presented to them (Bortolotti et al., 2012). This invariance means that the parameters behave similarly in different populations, allowing a uniform scale of measurement for use in different groups of individuals, and allowing individual ability levels to be assessed using different test forms or different subsets of test items (Magno, 2009). Overall, this enables IRT to infer much more about how individuals will respond to an item than other existing theories, such as CTT.

Additional studies highlight the benefits of IRT in facilitating meaningful scaling of item and individual parameters. IRT allows for greater accuracy than other techniques when scaling of individual and item parameters simultaneously, contributing to our understanding of the relationships between items and latent traits (Ryan & Brockmann, 2009). Separate item and

individual parameters also enable greater precision in identifying measurements of Differential Item Functioning, as legitimate differences in group abilities can be accurately distinguished from test bias, and test bias can itself be examined in greater detail (Thomas, 2011).

4. SMART Modules

4.1. Logical Reasoning

Every time a person discovers links between various elements, logical faculties are being utilized, which is the intellectual capacity most strongly related to general mental ability. In the professional environment, numerous activities rely on a highly developed sense of logic such as in law, information technology, medicine, etc.

There is substantial evidence that measures of reasoning and general mental ability are highly correlated with performance in the workplace (Schmidt & Hunter, 2004). One of the best and most well understood ways to assess general mental ability is through abstract reasoning tests, where individuals are given figural information and must decide on the answer that most accurately and consistently fits the pattern. One reason for the popularity and widespread use of these tests is that they are at best 'culture-free' (Cattell, 1971), and at worst are less confounded by cultural factors than other assessments of aptitude or ability which tend to assess domain specific types of ability (e.g., numerical or verbal reasoning).

Higher levels of fluid intelligence are associated with higher outcomes on:

- The identification of patterns and trends when given new information
- Thinking clearly and carefully about complex ideas
- Making logical conclusions about problems
- Learning new skills and understanding new information

Additionally, lower levels of fluid intelligence will mean an individual will likely have difficulty:

- Coming to correct, or logical conclusions
- Supporting decisions with evidence or coherent rationales
- Understanding complex relationships
- Identifying problems and understanding their causes

4.1.1. Development

In developing the **SMART – LOGICAL** reasoning items, the research and innovation team followed the guidelines laid out by Thompson & Weiss (2011). Broadly, these are:

- The development of an item bank
- Pretesting and calibration of items
- Integration within the CAT system

The first step required the development of a large volume of items measuring abstract reasoning. As can be seen in the example below (Figure 2), these items consist of 3 x 3 matrices, with each square barring the bottom right containing a pattern of shapes. The arrangement of shapes varies across each square according to a set of rules. When responding to an item, individuals are required to identify which of a number of patterns will best fit, given the information provided.

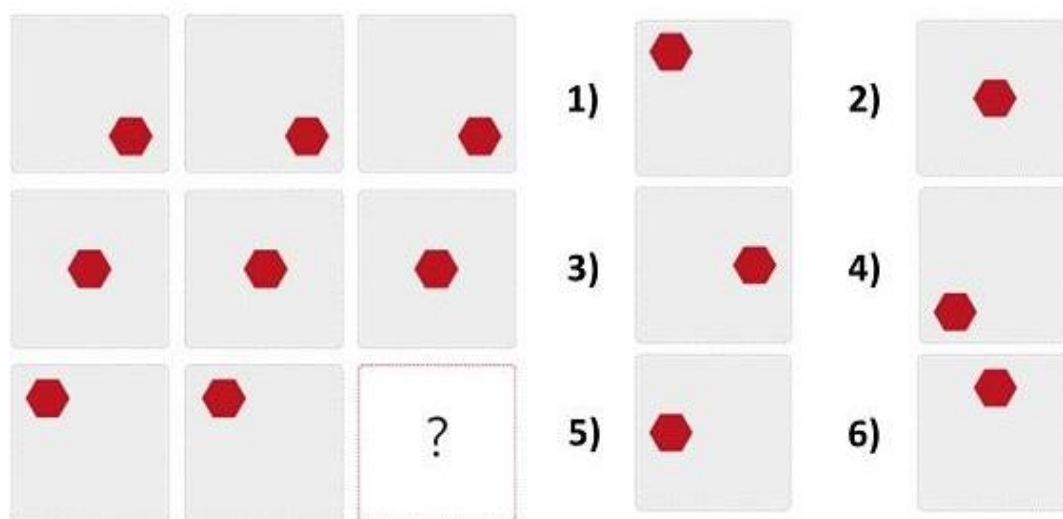


Figure 2. **SMART – LOGICAL** Example Item.

Once created, it was necessary to test the items in order to calculate the item parameters that would later allow selection in the adaptive test. Given the large volume of items, it was impractical to administer the entire bank to individuals. Instead, anchor items were selected and administered alongside one of a number of alternate tests. These anchor items were selected with an initial pretest, to determine which were of a 'medium' difficulty. Research indicates that selection of anchor items in this difficulty range is the most appropriate, when conducting linking studies (Sinharay, Haberman, Holland, and Lewis, 2012).

10 alternate forms of the assessment, plus anchor items, were administered to participants (N = 2137) of the *Research and Innovation Testing Center of Central Test*. Tests were administered with both French and English instructions, and differential item function analysis showed no difference in the responses of French-speaking or English-speaking participants. Because there was no difference in the way items were responded to, and because the items rely on non-verbal reasoning abilities, IRT parameters were calculated on the basis of both languages pooled together.

4.1.2. Norms

Norms help to compare an individual's score on a test with the scores of others who completed the same assessment, usually under similar conditions. These individuals who took the test at the same time are representative of the population for which an assessment is intended, and are therefore referred to as the norm group.

With norms, test developers and psychologists are able to establish comparisons between individuals and make more objective decisions. Norms also help HR Managers, recruiters and coaches to a) determine a candidate's suitability for a role relative to the job specification; b) evaluate a candidate's abilities across a range of desired competencies; c) Identify an employee's skill deficits and tailor training for development; d) establish objective metrics of performance. Norm groups, additionally, enable in attributing meaning to test scores and to quantify an individual's performance, level of an attribute or trait relative to his/her peers.

In order to provide normative data for **SMART – LOGICAL**, performance on the subset of items which were common across test administrations (anchor items) was analysed separately from the performance on the list-dependent unique items.

Due to the adaptive nature of **SMART – LOGICAL**, norms have been provided for the non-converted scores (ranging from -3 to 3) in Tables 3 to 5. Note that the non-converted scores are not displayed in the report, and are specified in the manual in order to visualise the differences in the estimated parameters, as mentioned in Chapter 4.1.4.

Table 3. SMART - LOGICAL Norms by Gender.

Gender	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
Male	327	-0.69	5.4	1.02
Female	337	-0.76	4.7	1.01

Table 4. **SMART - LOGICAL** Norms by Age Group.

Age Group	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
18-25	178	-0.61	5.7	1.04
26-33	170	-0.63	5.7	1.00
34-41	137	-0.79	4.9	0.97
42-29	79	-0.78	4.6	0.97
50+	100	-0.97	3.9	1.06

Table 5. **SMART - LOGICAL** Norms by Study Level.

Study level	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
Level 1: Master's Degree, PhD, Engineering or Business School, MBA	50	-0.55	6.3	1.01
Level 2: Bachelor's Degree, First Year Masters	133	-0.62	5.1	0.87
Level 3: 2 Year of Vocational or University Studies	53	-0.67	5.3	1.01
Level 4: High School Diploma	43	-0.91	4.5	1.17

4.1.3. Reliability

Reliability measures the extent to which an assessment is free from error and the degree to which items on a scale measure the same construct. A primary concern, when developing and using assessments, is the extent to which random measurement errors influence test performance. If a test is unreliable, it is impossible for it to be valid, as accuracy of results

(validity) depends on the consistency (reliability) with which the same results are produced. Reliability enables test users to be confident that the results accurately reflect the test taker's attitude, aptitude or behaviour in the workplace. In the context of a work-based assessment, reliability helps assure test users that the scores obtained reflect how a job candidate or employee is likely to behave in the workplace, interact with others, and work on specific tasks. If an assessment has high reliability, it also helps to guard against biases that may be caused from errors in human judgement during the recruitment and selection process.

4.1.3.a. Internal Reliability

Internal reliability measures the extent to which items on a scale are internally consistent and all measure the same construct. Internal reliability is measured by calculating Cronbach's Alpha (α) for each factor, which computes the average of all possible inter-item correlations without relying on a single indicator of reliability, which may contain large amounts of error. Cronbach's Alpha ranges from 0 to 1, with a minimum recommended value of .7 (1 being equal to perfect reliability (Nunnally, 1978). Values greater than .9 are considered excellent, of .7 to .9 are considered good, of .6 to .7 are acceptable, of .5 to .6 are poor, and values of less than .5 are unacceptable.

The reliability for **SMART – LOGICAL** was computed on the combined nationality sample (n = 664) and shows an overall reliability of .80, indicating that the scale has good internal consistency and is highly reliable.

4.1.4. Item Parameters

Due to the nature of measuring logical reasoning ability, item parameters for **SMART – LOGICAL** were calculated using a 1-parameter logistic model (1PL, or 'Rasch' model), meaning that items are characterised by their difficulty only. *Item difficulty* shows where items in **SMART – LOGICAL** function along the ability scale, between -3 and +3 (as mentioned in Chapter 3.5.1.a.). Lower scores indicate that an item is easy, whereas higher scores indicate difficult items.

Table 6. **SMART – LOGICAL** Item Parameter Estimates.

Question No.	Item Difficulty	Question No.	Item Difficulty	Question No.	Item Difficulty
1	-1.19	35	0.519510854	68	0.326047841
2	-0.75	36	1.028810313	69	-0.857268049

3	1.04	37	1.419273231	70	-0.812881657
4	1.96	38	1.113693556	71	-0.049435725
5	0.85	39	-1.551640279	72	0.190250789
6	-3.65	40	0.366721016	73	-0.129331229
7	-0.03	41	1.478691502	74	0.39442819
8	-1.09	42	-0.22512786	75	0.021582502
9	1.13	43	-0.321358501	76	1.379806079
10	0.81	44	1.26207295	77	0.65186926
11	-1.5	45	-0.959980025	78	1.770406324
12	0.35	46	-0.942483545	79	-3.558688655
13	1.04	47	0.973381028	80	-0.987533607
14	1.04	48	-0.426337381	81	0.10564456
16	-0.501700281	49	0.579710225	82	-0.016791395
17	-0.234173829	50	-0.870144022	83	0.770296885
18	-0.121985316	51	-1.185987525	84	0.980187093
19	2.121784928	52	0.744167217	85	0.735315183
20	0.99126992	53	-0.062988402	86	0.875241989
21	-0.208284172	54	-0.856106533	87	0.131880836
22	0.741003239	55	0.659942283	88	0.770296885
23	0.775522781	56	1.102123188	89	-2.754087094
24	0.174006803	57	0.603792327	90	0.135341465
25	-1.571650387	58	0.807335918	91	-0.269883272
26	-2.25877077	59	-0.625138854	92	-1.106760446
27	-0.540969812	60	1.411519718	93	0.135341465
28	0.573279458	61	-1.71061073	94	0.276289199
29	2.272509595	62	0.348428705	95	0.478901567
30	-0.039557641	63	-0.815376193	96	1.606483444
31	1.761812013	64	1.411519718	97	1.113166373
32	1.186116557	65	1.288424969	98	0.619849302

33	-2.035474765	66	0.986283313	99	1.452916318
34	-1.016875847	67	2.351515981		

4.2. Numerical Reasoning

There are few occupations in the world that are free from any kind of numeracy requirements. Where jobs differ, it is in the extent to which individuals are required to manipulate numbers, and the sophistication of the inferences they are expected to make from numerical data.

SMART – NUMERICAL is a test of reasoning ability, measuring a person's ability to manipulate and reason using numerical information. Rather than assessing whether an individual can perform numerical calculations in their head, it instead focuses on their ability to follow the appropriate steps when transforming numbers.

Higher levels of numerical intelligence are associated with:

- Accurate and efficient manipulation of numerical datasets
- Improved confidence when dealing with numbers and figures
- Greater identification of errors in mathematical thinking
- Understanding and interpreting numerical outcomes

Additionally, lower levels of numerical intelligence will mean an individual will likely have difficulty:

- Performing calculations and manipulating datasets
- Confidently approaching tasks that require data analysis or mathematical transformations
- Working accurately and efficiently with numerical data
- Understanding the implications of numerical outcomes

4.2.1. Development

In developing the **SMART – NUMERICAL** reasoning items, the research and innovation team followed the guidelines laid out by Thompson & Weiss (2011). Broadly, these are:

- The development of an item bank
- Pretesting and calibration of items
- Integration within the CAT system

The first step required the development of a large volume of items measuring numerical reasoning. As can be seen in the example below (Figure 3), these items consist of mathematical problems requiring the conversion of numbers from one scale to another (such as time and distance, currencies, etc.).

After ordering 200 business cards, Adrian noticed that 50 of them weren't printed properly. If he receives a refund for each business card that wasn't printed correctly, and he paid \$29 for the 200, how much should he expect back?
a) \$ 9.25
b) \$ 14.50
c) \$ 9.67
d) \$ 7.25

Figure 3. **SMART – NUMERICAL** Example Item.

Once created, it was necessary to test the items in order to calculate the item parameters that would later allow selection in the adaptive test. Given the large volume of items, it was impractical to administer the entire bank to individuals. Instead, anchor items were selected and administered alongside one of a number of alternate tests. These anchor items were selected with an initial pretest, to determine which were of a 'medium' difficulty. Research indicates that selection of anchor items in this difficulty range is the most appropriate, when conducting linking studies (Sinharay, Haberman, Holland, and Lewis, 2012).

Once the 'medium' difficulty level anchor items had been selected, 10 alternate forms of the assessment, plus anchor items, were administered to participants (N = 2019) of the *Research and Innovation Testing Center of Central Test*. The data from this was then used to calculate the IRT parameters of each item.

4.2.2. Norms

Norms help to compare an individual's score on a test with the scores of others who completed the same assessment, usually under similar conditions. These individuals who took the test at the same time are representative of the population for which an assessment is intended, and are therefore referred to as the norm group.

With norms, test developers and psychologists are able to establish comparisons between individuals and make more objective decisions. Norms also help HR Managers, recruiters and coaches to a) determine an individual's suitability for a role relative to the job specification; b)

evaluate an individual's abilities across a range of desired competencies; c) Identify an individual's skill deficits and tailor training for development; d) establish objective metrics of performance. Norm groups, additionally, enable in attributing meaning to test scores and to quantify an individual's performance, level of an attribute or trait relative to his/her peers.

In order to provide normative data for **SMART – NUMERICAL**, performance on the subset of items which were common across test administrations (anchor items) was analysed separately from performance on the list-dependent unique items.

Due to the adaptive nature of **SMART - NUMERICAL**, norms have been provided for the non-converted scores (ranging from -3 to 3) in Tables 7 to 9. Note that the non-converted scores are not displayed in the report, and are specified in the manual in order to visualise the differences in the estimated parameters, as mentioned in Chapter 4.2.4.

Table 7. SMART – NUMERICAL Norms by Gender.

Gender	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
Male	188	-0.83	5.0	1.16
Female	103	-0.78	5.0	1.20

Table 8. SMART – NUMERICAL Norms by Age Group.

Age Group	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
18-25	65	-0.79	5.7	1.18
26-33	67	-1.07	3.9	1.19
34-41	62	-0.81	4.8	1.10
42-29	41	-0.57	5.0	1.22
50+	56	-0.72	5.9	1.18

Table 9. **SMART – NUMERICAL** Norms by Study Level.

Study Level	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
Level 1: Master's Degree, PhD, Engineering or Business School, MBA	59	-0.62	5.5	1.25
Level 2: Bachelor's Degree, First Year Masters	85	-0.74	5.7	1.17
Level 3: 2 Year of Vocational or University Studies	47	-0.91	4.2	1.29
Level 4: High School Diploma	49	-0.91	4.0	0.98

4.2.3. Reliability

Reliability measures the extent to which an assessment is free from error and the degree to which items on a scale measure the same construct. A primary concern, when developing and using assessments, is the extent to which random measurement errors influence test performance. If a test is unreliable, it is impossible for it to be valid, as accuracy of results (validity) depends on the consistency (reliability) with which the same results are produced. Reliability enables test users to be confident that the results accurately reflect the test taker's attitude, aptitude or behaviour in the workplace. In the context of a work-based assessment, reliability helps assure test users that the scores obtained reflect how a job candidate or employee is likely to behave in the workplace, interact with others, and work on specific tasks. If an assessment has high reliability, it also helps to guard against biases that may be caused from errors in human judgement during the recruitment and selection process.

4.2.3.a. Internal Reliability

Internal reliability measures the extent to which items on a scale are internally consistent and all measure the same construct. Internal reliability is measured by calculating Cronbach's Alpha (α) for each factor, which computes the average of all possible inter-item correlations without relying on a single indicator of reliability, which may contain large amounts of error. Cronbach's Alpha ranges from 0 to 1, with a minimum recommended value of .7 (1 being equal to perfect reliability (Nunnally, 1978). Values greater than .9 are considered excellent, of .7 to .9 are

considered good, of .6 to .7 are acceptable, of .5 to .6 are poor, and values of less than .5 are unacceptable.

The reliability for **SMART – NUMERICAL** was computed on the English-speaking sample (n = 291) and shows an overall reliability of .79, indicating that the scale has good internal consistency and is highly reliable.

4.2.4. Item Parameters

Under the 2-parameter logistic model, item parameters represent item discrimination and difficulty. *Item difficulty* shows where items in **SMART – NUMERICAL** function along the ability scale, between -3 and +3. Lower scores indicate that an item is easy, whereas higher scores indicate difficult items. *Item discrimination* indicates how well an item is able to discriminate between the ability level of test-takers (as mentioned in Chapter 3.5.1.a).

Table 10. **SMART – NUMERICAL** Item Parameter Estimates.

Question No.	Item Discrimination	Item Difficulty	Question No.	Item Discrimination	Item Difficulty
1	1.200471	0.214463	44	1.316654	0.36535
2	0.354091	2.87443	45	0.544254	0.761825
3	0.693992	1.235417	47	1.079591	0.022705
5	1.394108	-1.66315	49	1.226947	0.549937
6	1.059298	-1.69422	51	0.720241	-1.84896
7	1.733555	-2.16342	52	1.220653	0.180278
8	1.662592	-0.94022	53	0.862049	2.531468
9	1.123863	-0.68811	54	1.998005	-0.98709
10	2.18143	-2.75465	55	1.986283	0.425186
11	1.997683	-1.92273	56	0.770447	-0.08562
12	0.418602	-0.58981	57	0.979343	1.094297
13	1.086932	-0.60662	58	1.760015	0.668321

15	1.26434	0.542479	59	1.793653	0.223925
16	2.084487	-1.06949	60	0.638254	0.551862
17	0.842635	0.276327	62	1.138285	1.764006
18	1.653579	-1.24342	63	0.889515	0.276086
19	1.230073	-0.42784	64	0.622503	3.068096
20	1.061408	-0.14326	66	1.614724	-0.36476
21	1.182653	1.314226	67	1.238487	-0.60963
22	1.253577	0.661653	68	1.507086	0.951984
23	0.499678	1.451504	69	0.617735	2.462699
24	1.796099	-1.00893	70	2.383153	-0.18451
25	0.480533	1.459023	72	0.935947	0.566906
26	2.020719	-0.32026	74	1.419279	-0.63347
27	0.830353	-1.10593	75	0.979585	0.005501
29	1.728605	-0.32803	76	1.082823	-0.0863
30	0.710177	0.317925	77	0.861567	0.639008
32	1.3926	-0.62914	78	1.523215	-0.92182
33	1.132831	-0.03957	80	1.362979	-0.71475
34	0.979578	1.174983	81	1.058308	-0.2451
35	0.842863	0.383319	82	1.95973	0.719501
36	1.118212	1.33636	83	0.456943	0.383835
37	1.768498	-0.71329	84	1.068589	0.499846
38	1.743354	-0.58681	85	1.422882	-0.06074
39	1.373615	0.849993	86	0.509093	0.688041
40	1.391182	0.186389	87	0.870592	0.980272
41	0.646444	1.690537	88	1.757727	-0.4915
43	0.425898	1.109006	89	1.792078	-0.87805

4.3. Verbal Reasoning

SMART – VERBAL is a test of cognitive ability, measuring a person's ability to manipulate and reason using verbal information. Specifically, the assessment tests "analogical" reasoning; the ability to reason by way of analogies. Analogies require an individual to understand logical relationships between different concepts and meanings, understanding how things might be considered similar or dissimilar.

Higher levels of verbal reasoning are associated with:

- Better communication skills, and improved efficiency when convincing or persuading others
- Improved confidence when dealing with complex arguments, or learning new concepts
- Greater identification of errors in verbal reasoning, or in arguments
- Understanding and interpreting the outcomes of verbally presented arguments

Additionally, lower levels of verbal intelligence will mean an individual will likely have difficulty:

1. Communicating effectively and efficiently, or convincing and persuading others
2. Confidently approaching tasks that require understanding of complex arguments, or learning new concepts
3. Working accurately and efficiently with verbal information
4. Understanding the implications of verbally presented arguments

4.3.1. Development

In developing the **SMART – VERBAL** reasoning items, the research and innovation team followed the guidelines laid out by Thompson & Weiss (2011). Broadly, these are:

- The development of an item bank
- Pretesting and calibration of items
- Integration within the CAT system

The first step required the development of a large volume of items measuring verbal reasoning. As can be seen in the example below (Figure 4), these items consist of analogies requiring the linking of elements from a verbal structure to another element from another verbal structure while applying two principles.

(A): Tree, Jungle, Leaf (B): Nature, Cow, Grass

(A) is to forest as (B) is to meadow.

Figure 4. **SMART – VERBAL** Example Item.

Once created, it was necessary to test the items in order to calculate the item parameters that would later allow selection in the adaptive test. Given the large volume of items, it was impractical to administer the entire bank to individuals. Instead, anchor items were selected and administered alongside one of a number of alternate tests. These anchor items were selected with an initial pretest, to determine which were of a 'medium' difficulty. Research indicates that selection of anchor items in this difficulty range is the most appropriate, when conducting linking studies (Sinharay, Haberman, Holland, and Lewis, 2012).

Once the 'medium' difficulty level anchor items had been selected, 10 alternate forms of the assessment, plus anchor items, were administered to participants (N = 767) of the *Research and Innovation Testing Center of Central Test*. The data from this was then used to calculate the IRT parameters of each item.

4.3.2. Norms

Norms help to compare an individual's score on a test with the scores of others who completed the same assessment, usually under similar conditions. These individuals who took the test at the same time are representative of the population for which an assessment is intended, and are therefore referred to as the norm group.

With norms, test developers and psychologists are able to establish comparisons between individuals and make more objective decisions. Norms also help HR Managers, recruiters and coaches to a) determine an individual's suitability for a role relative to the job specification; b) evaluate an individual's abilities across a range of desired competencies; c) Identify an individual's skill deficits and tailor training for development; d) establish objective metrics of performance. Norm groups, additionally, enable in attributing meaning to test scores and to quantify an individual's performance, level of an attribute or trait relative to his/her peers.

In order to provide normative data for **SMART – VERBAL**, performance on the subset of items which were common across test administrations (anchor items) was analysed separately from performance on the list-dependent unique items.

Due to the adaptive nature of **SMART - VERBAL**, norms have been provided for the non-converted scores (ranging from -3 to 3) in Tables 11 to 13. Note that the non-converted scores are not displayed in the report, and are specified in the manual in order to visualise the differences in the estimated parameters, as mentioned in Chapter 4.3.4.

Table 11. SMART – VERBAL Norms by Gender.

Gender	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
Male	170	-0.80	4.8	0.94
Female	110	-0.61	5.8	1.01

Table 12. SMART – VERBAL Norms by Age Group.

Age Group	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
18-25	69	-0.68	5.4	1.03
26-33	66	-0.68	5.2	0.92
34-41	63	-0.84	4.4	0.99
42-29	38	-0.80	4.7	1.02
50+	44	-0.65	5.8	0.91

Table 13. SMART – VERBAL Norms by Study Level.

Study Level	Sample Size	Mean Non-converted Score	Mean Score	Standard Deviation
Level 1: Master's Degree, PhD, Engineering or Business School, MBA	78	-0.89	4.3	1.06

Level 2: Bachelor's Degree, First Year Masters	82	-0.69	5.3	1.05
Level 3: 2 Year of Vocational or University Studies	35	-0.61	6.3	0.71
Level 4: High School Diploma	54	-0.73	4.7	0.94

4.3.3. Reliability

Reliability measures the extent to which an assessment is free from error and the degree to which items on a scale measure the same construct. A primary concern, when developing and using assessments, is the extent to which random measurement errors influence test performance. If a test is unreliable, it is impossible for it to be valid, as accuracy of results (validity) depends on the consistency (reliability) with which the same results are produced. Reliability enables test users to be confident that the results accurately reflect the test taker's attitude, aptitude or behaviour in the workplace. In the context of a work-based assessment, reliability helps assure test users that the scores obtained reflect how a job candidate or employee is likely to behave in the workplace, interact with others, and work on specific tasks. If an assessment has high reliability, it also helps to guard against biases that may be caused from errors in human judgement during the recruitment and selection process.

4.3.3.a. Internal Reliability

Internal reliability measures the extent to which items on a scale are internally consistent and all measure the same construct. Internal reliability is measured by calculating Cronbach's Alpha (α) for each factor, which computes the average of all possible inter-item correlations without relying on a single indicator of reliability, which may contain large amounts of error. Cronbach's Alpha ranges from 0 to 1, with a minimum recommended value of .7 (1 being equal to perfect reliability (Nunnally, 1978). Values greater than .9 are considered excellent, of .7 to .9 are considered good, of .6 to .7 are acceptable, of .5 to .6 are poor, and values of less than .5 are unacceptable.

The reliability for **SMART – VERBAL** was computed on the English-speaking sample (n = 280) and shows an overall reliability of .82, indicating that the scale has good internal consistency and is highly reliable.

4.3.4. Item Parameters

Under the 2-parameter logistic model, item parameters represent item discrimination and difficulty. *Item difficulty* shows where items in **SMART – VERBAL** function along the ability scale, between -3 and +3. Lower scores indicate that an item is easy, whereas higher scores indicate difficult items. *Item discrimination* indicates how well an item is able to discriminate between the ability level of test-takers (as mentioned in Chapter 3.5.1.a).

Table 14. **SMART – VERBAL** Item Parameter Estimates.

Question No.	Item Discrimination	Item Difficulty	Question No.	Item Discrimination	Item Difficulty
1	0.856104	-0.74947	51	0.359504	2.380002
2	0.552158	-1.38944	53	1.021844	1.717866
3	1.301293	-0.21969	54	0.95917	1.220956
4	1.376956	0.389095	55	0.773448	0.541722
5	0.709068	0.518034	57	0.494692	-2.17416
6	10.77934	2.134362	58	1.117098	-1.067
7	0.890321	0.202534	59	0.664773	1.095343
8	1.404458	-1.09106	60	0.539173	-2.80538
9	1.023461	1.778371	61	1.85607	-1.24764
10	3.274149	-0.3218	62	0.461654	-0.13135
11	1.106733	-0.46481	63	0.637701	1.329665
12	0.965041	-0.85609	64	1.298989	1.735424
14	1.031823	0.251699	65	0.107562	1.708157
15	0.616283	0.218774	66	0.79873	-1.11259
16	0.477142	1.440151	67	0.921114	-1.51374
18	0.268736	-3.9557	68	0.964414	-1.60946
19	0.755506	-2.53705	69	0.792812	-0.46172

20	1.953398	-0.73336	70	0.21117	4.673374
21	1.671175	-0.46674	71	0.782622	-0.39793
22	0.171734	-2.29967	73	1.524534	0.829154
23	1.391845	0.725494	74	0.875815	4.816588
25	1.188797	-0.7852	75	0.641753	-2.52472
26	1.215909	1.969175	76	2.383728	-1.58062
27	0.257679	-2.8613	77	1.199127	-1.83997
28	0.92348	-0.25432	78	1.022497	-0.9071
29	1.002359	-4.16757	79	0.725068	-2.82373
30	1.189274	0.067604	80	0.941696	-3.90181
31	0.753486	-0.1727	81	1.060643	3.149496
33	0.559111	0.440068	82	0.146594	7.828225
35	0.788416	1.427027	83	0.968102	1.813545
38	1.105152	-0.69447	84	1.29772	-0.47249
39	1.445378	0.76329	85	1.258356	0.403911
40	2.030308	0.780229	86	1.238385	2.377037
41	0.352807	-1.20067	87	0.589477	-0.96562
43	1.710841	2.024827	88	0.634346	1.774736
44	0.715848	0.523983	89	0.493801	-1.8028
45	0.491333	4.167154	90	1.109319	1.475745
46	0.786385	3.734811	93	0.592519	-1.42972
47	0.713264	-2.60667	94	0.80641	-0.95168
49	0.647209	-1.43073	96	1.487522	1.780015
50	0.494163	0.801504	97	0.778349	0.517813

5. Scoring

5.1. Percentile Rank Scores and the Scoring Method

In order to provide an easily interpretable score that enables an efficient indication of where an individual's score lies in comparison to the normed population, scores on the **SMART** reports are generated as percentile rank scores.

Individuals receive a score between 0 - 10, indicating their percentile rank and denoting how the individual has performed in relation to the scores of the comparison group, thereby allowing meaningful comparisons to be made. This is visually represented on a graph displaying the individual's percentile rank as a percentage between 0 - 100. For example, a score of 7.4 equates to a percentile rank of 74%. A percentile rank of 74% indicates that the individual has scored higher than 74% of the comparison group; it does not mean that the individual has scored 74% in the assessment. The comments in each **SMART** module follow this scoring methodology, while providing an extensive interpretation of the individual's reasoning skill across 4 levels of ability. For example, lower percentile rank scores will be accompanied by comments indicating that the individual may have difficulty performing tasks which require the given reasoning skill from the **SMART** module, while higher percentile rank scores are accompanied by comments indicating that the individual has an advanced level of the reasoning ability.

5.2. Combined Score

When an individual takes all three **SMART** adaptive modules, they receive a combined score on the *Central Test* Talent Insights Report, also known as a general intelligence score. To calculate this combined score, a weighting of each **SMART** module is calculated. These are estimated through a factor analysis, which is conducted on a sample of participants who have taken all modules. This factor analysis enables us to determine the *g* loading of each module, i.e., how much each of the reasoning modules correlates with, or loads onto, the *g* factor of intelligence.

In the case where there is not enough data to run a factor analysis, literature can be used to determine the *g* loading of each module.

Table 15. Current *g* Factor Loadings, for Each of the Reasoning Skills Measured by the **SMART** Adaptive Modules.

Reasoning Skill	<i>g</i> loading
Numerical	0.44
Verbal	0.56
Logical	0.71

References

- Baker, F. (2001). *The Basics of Item Response Theory (2nd ed.)*. College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.
- Bortolotti, S., Tezza, R., de Andrade, D., Tezza, R., & de Sousa Junior, A. (2012). Relevance and advantages of using the item response theory. *Qual Quant* 47, 2341–2360.
- Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54(1), 1-22.
<http://dx.doi.org/10.1037/h0046743>
- Cattell, R. B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.
- Cattell, R. B. (1987). *Intelligence: Its Structure, Growth, and Action*. Elsevier Science Publishers.
- Chae, Sunhee, Kang, Unson, Jeon, Eunhwa, & Linacre, J. M. (2000). Development of Computerized Middle School Achievement Test. *MESA Memorandum No. 69*. Seoul, South Korea: Komesa Press.
- Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Orlando, FL: Harcourt.
- Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.) *Computerized adaptive testing: A Primer*. Mahwah, NJ: Erlbaum.
- Galton, F. (1869). *Hereditary Genius*. London: MacMillan.
- Gade, P. A., Johnson, R. M., & Pliske, R. M. (1984). Cross-Validation of the Computerized Adaptive Screening Test (CAST). *Research Report 1372*. U. S. Army Research Institute for the Behavioral and Social Sciences.
- Hally, T. (2012). *The g-factor*. https://www.researchgate.net/publication/275659645_The_g-Factor
- Kolen, M. F., & Brennan, R. L. (2014). *Test Equating, Scaling and Linking*. New York, NY: Springer.
- Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *International Journal of Educational and Psychological Assessment*, 1(1), 1-11.

- Meijer, R. R. (1999). Computerized Adaptive Testing: Overview and Introduction. *Applied Psychological Measurement*, 23 (3), 187-194.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Rocklin, T. R., Roos, L. L., Wise, S. L., & Yoes, M. E. (1996). Conducting Self-Adapted Testing Using MicroCAT. *Educational and Psychological Measurement*, 56 (5), 821-827.
- Ryan, J., & Brockmann, F. (2009). A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory. Retrieved from <https://eric.ed.gov/?id=ED544690>
- 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), 162-173.
- Sinharay, S., Haberman, S., Holland, P., & Lewis, C. (2012). A Note on the Choice of an Anchor Test in Equating. ETS research report series.
- Spearman, C. (1927). *The Abilities of Man*. London: Macmillan.
- Thomas, M. (2011). The Value of Item Response Theory in Clinical Assessment: A Review. *Assessment*, 18(3), 291-307.
- Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*. 16 (1)
- Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.
- Vinney, C. (2020). Fluid Versus Crystallized Intelligence: What's the Difference? <https://www.thoughtco.com/fluid-crystallized-intelligence-4172807>
- Vispoel, W., Rocklin, T., Wang, T., & Bleiler, T. (1999). Can Examinees Use a Review Option to Obtain Positively Biased Ability Estimates on a Computerized Adaptive Test? *Journal of Educational Measurement*, 36(2), 141-157.
- Wechsler, D. (1944). *The measurement of adult intelligence (3rd ed.)*. Baltimore: Williams & Wilkins.
- Weiss, D. J. (1983). *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing*.

CONTACT US

Client Service: customer@centraltest.com

Training and Consulting: consulting@centraltest.com



www.centraltest.com

*Central Test – Empowering Talent © 2021
All rights reserved.*