PSYTECH
INTERNATIONAL

Adaptive General Reasoning Test (AdaptGRT):

Working Technical Document

Prepared by Dr Jurgen Becker on behalf of Psytech International.

TABLE OF CONTENTS

TABLE OF FIGURES

1. INTRODUCTION TO COGNITIVE ABILITY TESTING

General intelligence, or cognitive ability as it is sometimes called, has been studied for more than a century (Spearman, 1904) due to its critical importance in education and the workplace. Kuncel, Hezlett, and Ones (2001), for example, showed that cognitive ability is an excellent predictor of academic performance as indexed by grades, degree completion, teacher ratings of performance, etc. For jobs of medium complexity, Hunter (1980) found that cognitive ability is strongly related to performance and the relation is even stronger for more complex jobs. Recently, Kuncel, Hezlett, and Ones (2004) demonstrated that a single measure of cognitive ability predicts a wide array of important outcomes including academic performance, career potential, creativity, and job performance. Based on data from thousands of studies, the inescapable conclusion is that cognitive ability is strongly related to performance in many important domains.

Carroll's (1993) three-stratum model is one of the best known and most widely accepted conceptualisation of intelligence that is currently available. The model separates the broad construct of cognitive ability into three distinct layers or strata that range from narrow, specific factors at Stratum I to a broad conceptualisation of 'g' or general intelligence at Stratum III. Stratum II consists of eight broad cognitive abilities that have an influence on g, including Gf and Gc.

Gf is Cattell's (1971) "fluid intelligence", which he defined as the ability to "educe complex relations among simple fundaments whose properties are known to everyone" (p. 98). Put more simply, Gf refers to basic reasoning ability as exemplified in inductive and deductive reasoning. Gc is "crystallised intelligence" and reflects an individual's acquired knowledge. As an example, word knowledge is a Stratum I ability strongly related to Gc.

In more recent years, the similarities of both Cattell's (1971) and Carroll's (1993) theories, as well as the work of Horn (1968) in association with Cattell, have been synthesised into an overarching theory termed the CHC (Cattell-Horn-Carroll) model. The consensus of many modern theorists is that the CHC taxonomy best explains the structure of cognitive ability and should be used to establish a nomenclature in the field of human intelligence (McGrew, 2009).

The CHC model also recognises a three-stratum structure that is similar to Carroll's model. General intelligence or 'g' falls within Stratum III, followed by a range of broad cognitive abilities in Stratum II (Schneider & McGrew, 2012). Within Stratum I falls the specific narrow abilities, such as spelling ability and writing ability. The broad cognitive domains outlined in Stratum II are:

- G*s*: Processing speed
- G*f:* Fluid reasoning
- G*c:* Comprehension-knowledge (crystallised intelligence)
- G*wm:* Working memory
- G*lr:* Long term storage and retrieval of information
- G*v:* Visual processing
- G*a:* Auditory processing

In line with the commonly accepted CHC taxonomy of human intelligence, a good measure of cognitive ability should incorporate assessments of both Gf and Gc. The Adaptive General Reasoning Test (AdaptGRT) was designed with this in mind. It contains two tests (Numerical Reasoning and Abstract Reasoning) that tap fluid intelligence and one test (Verbal Reasoning) that is a good measure of crystallised intelligence. The overall test score, which is a composite of Numerical Reasoning, Abstract Reasoning and Verbal Reasoning, provides an excellent measure of general cognitive ability.

2. COMPUTERIZED ADAPTIVE TESTING AND ITEM RESPONSE THEORY

Computerized Adaptive Testing, or CAT, is an onscreen form of testing that adapts to an individual's level of ability (Van der Linden & Glas 2000). In other words, the test adjusts the difficulty of the presented items based on an examinee's sequence of responses. If an individual answers an item of medium difficulty incorrectly, a CAT will present an easier item. An individual who correctly responds to an item will receive a more difficult question. Termination criteria build into the CAT engine determine when the optimal level of information has been achieved.

2.1 PRINCIPLES OF CAT

According to Weiss & Kingsbury (1984), there are five aspects that make up a CAT test, namely the item pool, the starting point of the test, item selection, termination criteria and the scoring procedure. These are converted into algorithms that regulate the test session and are contained in what is called the "engine" or software of the test.

2.1.1 Item pool

A CAT assessment requires a large pool of items. These items must be validated and calibrated before they are incorporated into the CAT in order to determine their difficulty level and other important parameters. These parameter estimates are introduced into the CAT engine and guide the selection of items. Items are calibrated according to Item Response Theory (IRT) – the statistical model that underlies CAT. It is important that a sufficient number of items are included in the total item pool in order to cover all levels of difficulty and examinee ability.

2.1.2 Starting Point

Initially an examinee's ability level is unknown. Certain CAT engines assume that he or she has an average level of ability and therefore present an item of medium difficulty at the outset of the test. Alternatively, as in the AdaptGRT, the examinee is presented with a number of items that range across the difficulty scale in order to determine a preliminary estimation of ability.

2.1.3 Item Selection

An algorithm guides the selection of items in the CAT in order to provide a maximal ability estimation with the fewest amount of items (Thissen & Mislevy, 2000). Once the starting point is established, a successive item that provides the greatest amount of information is selected. The initial ability estimation is continually updated after every correct or incorrect response and guides the selection of the following item.

2.1.4 Termination of the test

An algorithm is incorporated into the CAT engine to prevent item selection from continuing indefinitely. This criterion specifies the point at which the test should terminate. In the AdaptGRT test, this occurs when the standard error of measurement falls below a pre-specified threshold.

2.1.5 Scoring

As with item calibration, the scoring procedure that underlies ability estimation is based on IRT. More information regarding the IRT scoring model is provided in further on in chapter two.

2.2 ADVANTAGES AND LIMITATIONS OF CAT

CAT methodology offers a number of advantages over classical paper-and-pencil tests (Unick, Shumway & Hargreaves, 2008). Examinees no longer need to respond to a fixed number of items. Instead, items are maximally selected to provide the greatest amount of information into each examinee's level of underlying ability. Fewer test questions within a shorter testing session are required to obtain an accurate test score. This is beneficial to the candidate who is able to avoid fatigue associated with a long examination period, and to the organisation who is able to assess more people in the same span of time.

Each examinee also receives a unique version of the test as the sequence of items is determined based on his or her previous responses. This implies greater test security and less concern around memory effects in instances where the test is repeated.

However, as with all assessment methodologies, CAT may fall prey to a number of shortcomings (Meijer & Nering, 1999):

- The item selection algorithm implies the need for an item pool that can provide information at all levels of the difficulty spectrum. This pool of items is significantly larger than that required in traditional paper-and-pencil tests.
- Each test item must be administered within a pilot study and calibrated in order to be included in the CAT engine. Item calibration requires a large sample size in order to obtain accurate parameter estimates.
- As the majority of examinees possess levels of ability within the average range, certain items that provide optimal information at that point are overexposed and are continually selected in testing sessions. If the item pool is not sufficiently broad, there may also be gaps in the ability continuum that result in certain items being overexposed. Such incidences can be reduced by increasing the number of items at all levels of the ability range and including an exposure control algorithm in the CAT engine.
- As the number of test items are not fixed, it can be difficult for candidates to manage their responses within the given time limit.
- It is time intensive and expensive to trial and update extensive item pools.

2.3 INTRODUCTION TO ITEM RESPONSE THEORY (IRT)

The AdaptGRT is based on the three-parameter logistic IRT model (3PLM; Birnbaum, 1968), which is well-suited for the types of cognitive ability items used on this test. In a nutshell, AdaptGRT tailors test difficulty to the ability level of each examinee by selecting items that meet content and "information" specifications. In this way, each test is individually designed to provide high accuracy and precision with far fewer items than would be required for a typical paper-and-pencil (non-adaptive) test.

Below we provide an overview of the 3PLM, introduce the concepts of item and test information, describe how trait scores and their standard errors are estimated, and provide details about the item selection algorithm underlying the AdaptGRT.

2.3.1 What is IRT?

IRT is a model-based statistical methodology that, similarly to classical test theory (CTT), aims to measure an individual's performance on a latent trait or construct. IRT is based on a number of assumptions regarding the mathematical relationship between traits and examinee responses. The term 'ability', denoted with the Greek letter theta (Θ), is used within IRT to refer to an individual's score or 'level' of the latent trait in question.

IRT assesses the probability of correctly responding to an item as a function of both the latent trait and certain item parameters. This same methodology is used to determine the likelihood of an examinee's ability as a function of observed item responses and, once again, item parameters.

Unlike CTT, which is highly dependent on norms, IRT item parameter calibration is person- or sample-free, while the estimation of examinee ability is item- or test-free (Van Der Linden & Hambleton, 1997):

- Sample-free or person-invariant implies that item parameters are not influenced by the usual sample considerations that affect CTT.
- Item-free or item-invariant implies that one candidate's ability estimation can be meaningfully compared to another, even if they did not respond to the same set of items.

2.3.2 The 3PLM IRT Model

One of the most parsimonious and well recognised IRT models is the one-parameter logistic or Rasch model (Rasch, 1960) for dichotomous data (correct-incorrect; agree-disagree). The Rasch model uses examinee trait level and just one item parameter, difficulty (the location of an item on the trait continuum), to predict the probability of answering an item correctly. In most psychological domains, however, more item properties (parameters) are used to model the probability of correct responses. Specifically, for multiple-choice items, like those often used in educational assessments, item discrimination (how well an item differentiates between examinees of different abilities) and guessing are also taken into account.

In the three-parameter logistic model (3PLM), each item is characterised by an item difficulty parameter (b), an item discrimination parameter (a) and a pseudo-guessing parameter (c). The c parameter represents the lower asymptote at which examinees with the lowest level of ability can provide the right answer due to guessing. The probability of a correct or positive response to item i for the 3PLM is given by:

(1)

$$P(u_i = 1|\Theta) = c_i + \frac{1 - c_i}{1 + \exp[-1.7a_i(\Theta - b_i)]}$$

Whereby:

- u_i denotes an examinee's response to item i. u_i = 1 if correct and 0 if incorrect.
- $P(u_i = 1|\Theta)$ is the probability of a correct response for a randomly chosen examinee having trait level Θ.
- 1.7 is a scaling factor that is included for historical reasons.

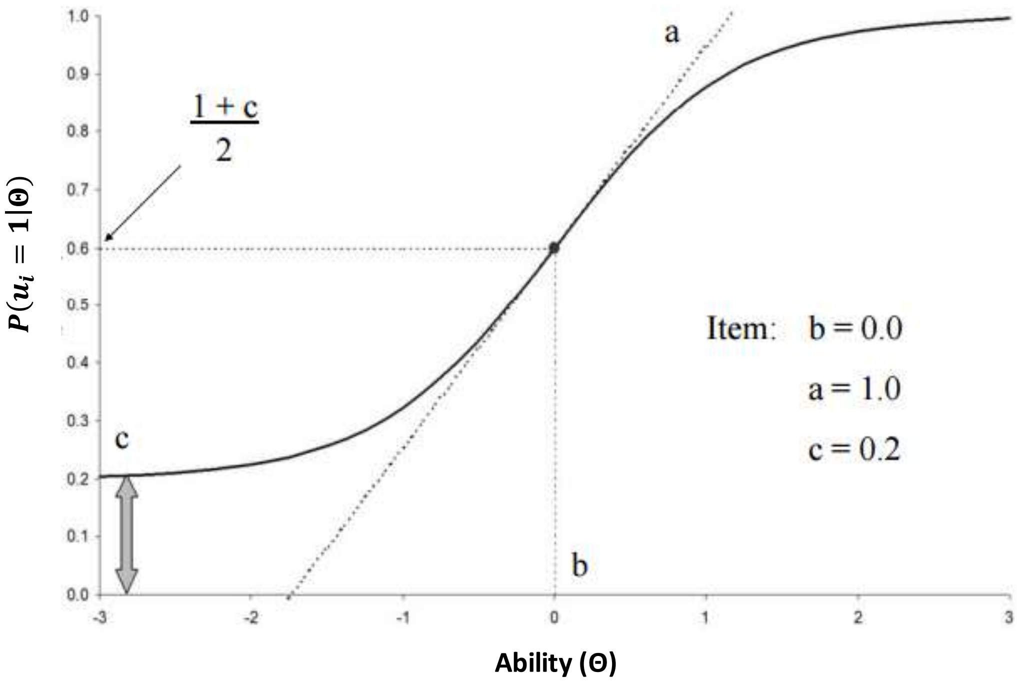
Note that the two-parameter logistic model (2PLM) and the one-parameter logistic model (1PLM or Rasch model), which are also sometimes used for cognitive ability testing, can be obtained from the equation above by placing "constraints" on the a- and c- parameters. If one assumes no guessing and sets c = 0 for all items, then the 2PLM results. If one also assumes that all items are equally discriminating (e.g., set all a = 1), then the 1PLM results.

The 3PLM is the IRT model that is most commonly applied in large scale cognitive ability testing programs. The presence of three item parameters makes this model flexible enough to describe responses to a wide variety of multiple-choice items. More importantly, numerous studies have shown that the 3PLM provides an excellent fit to cognitive ability data (e.g., Drasgow, Levine, Tsien, Williams, & Mead, 1995), so it can be readily used for scoring examinees.

2.3.3 Item response function (IRF)

The item response function (IRF) or item characteristic curve (ICC) represents the probability of answering an item correctly as a function of an examinee's ability level. Figure 1 presents an example of a 3PLM IRF with a difficulty (b) parameter of 0, a discrimination (a) parameter of 1 and a guessing (c) parameter of 0.2.

Figure 1: Item response function (IRF) demonstrating the parameters of 3PLM.



The x-axis of an IRF plot represents the theoretical ability of an examinee and typically ranges from -3 to +3. Note, the -3 to +3 range is chosen for Θ as its distribution is usually assumed to be standard normal during the initial phase of item parameter estimation and scoring. The y-axis represents the probability of obtaining a correct response, denoted by $P(u_i = 1|\Theta)$.

Figure 2 shows 3PLM IRFs for three hypothetical items having the same discrimination and guessing parameters (a = 1.5 and c = 0.2, respectively), but different difficulty (or "location") parameters (b = -1.0, 0.0, +1.0, respectively). It can be seen that the item difficulty parameter affects the lateral position of the IRFs along the trait continuum. As the difficulty parameter increases from -1.0 to 1.0, the probability of a correct response, at a particular theta, decreases. For example, only examinees having Θ >1.5 have a high probability of answering the last item correctly.

Figure 3 illustrates how the item discrimination parameter affects the shape of IRFs. Shown are items that exhibit rather low discrimination (a = 0.5), medium discrimination (a = 1.0), and high discrimination (a = 2.0). It is evident that as the item discrimination parameter increases, the IRFs become steeper in slope. Note that the difference in response probabilities for examinees with, say, Θ = -0.5 and Θ = 0.5 increases as item discrimination increases. For the low discrimination item, the response probability difference is only 0.2. However, for the high discrimination item, the difference in response probabilities is nearly 0.8. Thus, items with large a parameters better differentiate examinees having different trait levels; hence the term "item discrimination."

Figure 2: IRF plots for 3PLM – difficulty.

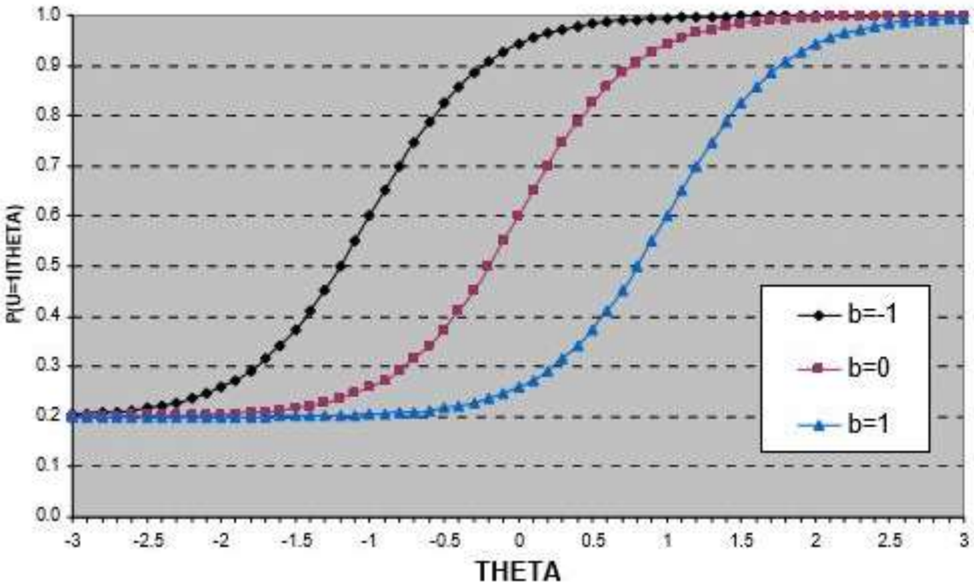
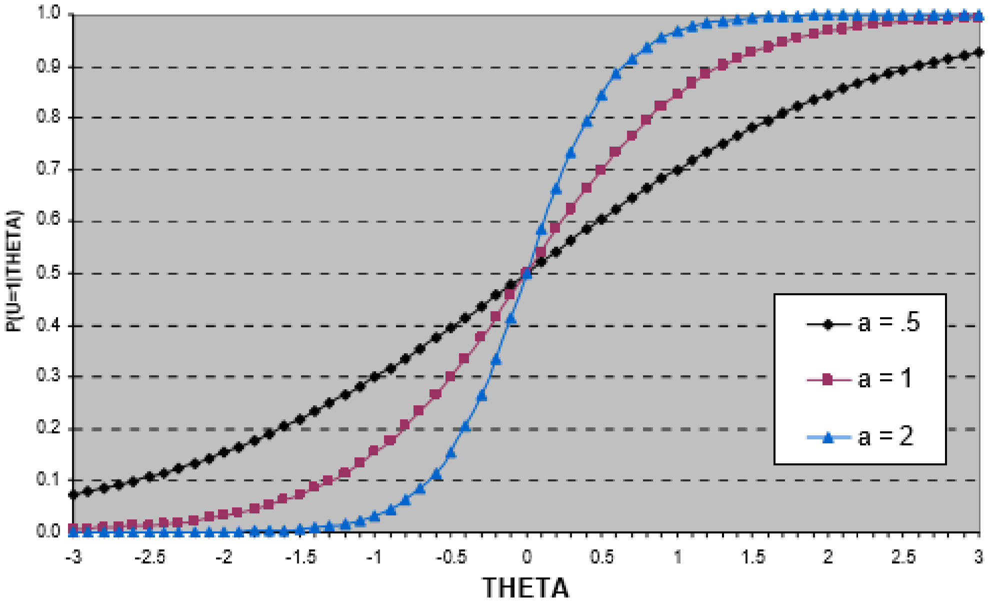
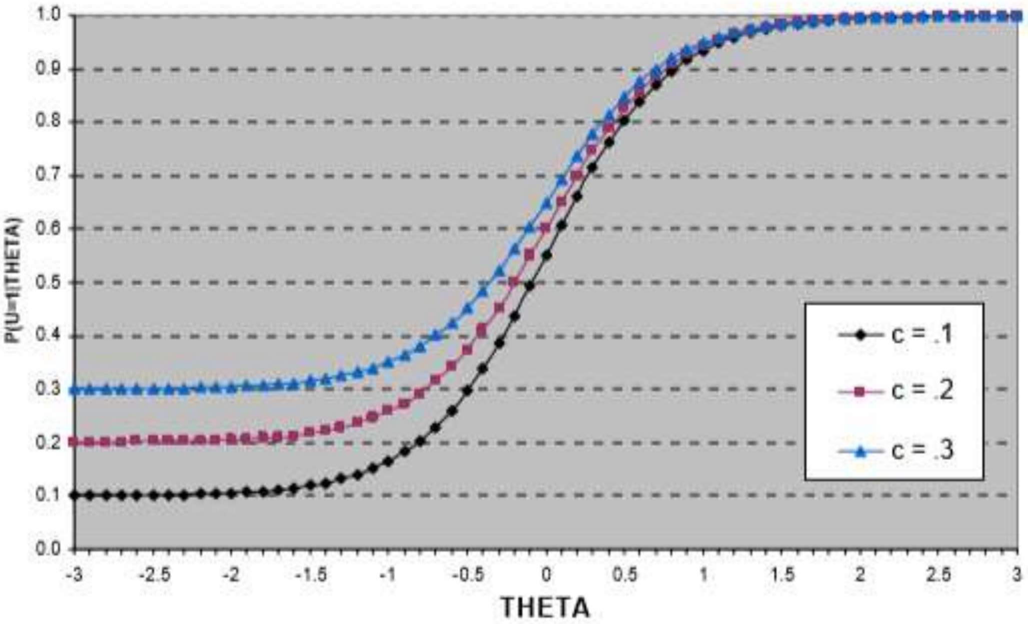


Figure 3: IRF plots for 3PLM – discrimination.



Finally, Figure 4 illustrates the effect of the third item parameter of the 3PLM, the guessing parameter, c. With a typical multiple-choice item, even low ability examinees can guess the correct answer. As shown, the c-parameter elevates the lower asymptote of the IRF so that the probability of a correct response remains above zero for any trait level. Values of the c- parameter typically range from 0.1 to 0.3 for items having 4 or 5 response options.

Figure 4: IRF plots for 3PLM – guessing.



2.3.4 Estimating Trait Levels Using Item Response Theory

The logic of scoring examinees in IRT is fundamentally different from Classical Test Theory (CTT) scoring. In CTT, trait levels are estimated by summing responses across items into a total score and then converting the summed score to a standard score or a score on another metric (e.g., the IQ metric with mean 100 and standard deviation 15 or the SAT metric with mean 500 and standard deviation 100). Summed scores are not appropriate for adaptive tests because some examinees are administered items that are more difficult than the items administered to other examinees. In contrast to CTT scoring, IRT scoring estimates an individual's trait level, Θ, which does not depend on the set of items that are administered.

In IRT, estimating trait levels is similar to clinical diagnosis, where a clinician is trying to estimate the most likely disease given a set of presenting symptoms. In IRT, the "symptoms" are item responses and the "disease" is an examinee's trait level. Note that both IRT and clinical diagnosis assume that other outcomes are also possible (an examinee may have a different trait level or a patient can have a different disease), but the diagnosed outcome (trait level) is the most likely one. Therefore, in IRT there is a search process in which the presenting behaviours (item responses and their parameters) are used to determine what trait level is most likely.

Consider for example an examinee with a correct and an incorrect response to two items. Suppose the correctly answered item has an item discrimination parameter of 1.0, an item difficulty parameter of -1.0 and a guessing parameter of 0.1. Suppose the incorrectly answered item has item parameters of 1.0, 2.0, and 0.2, respectively. The conditional probability of observing a response pattern (correct, incorrect), given the item parameters specified above, is, simply, a product of individual item response probabilities provided by Equation 1 above.

This product rule is due to IRT's assumption of local or conditional independence, which states that an examinee's item responses are independent of one another, given his or her Θ level. In essence, conditional independence states that the response probability for a given item is a function of that examinee's trait level and the item parameters, and does not depend on how the examinee responded to other items.

Formally, the conditional probability of a response pattern u = $< u_1, u_2, \dots, u_n >$ for an examinee j given Θ and item parameter vectors $\beta_1 =< a_1, b_1, c_1 >$ is given by:

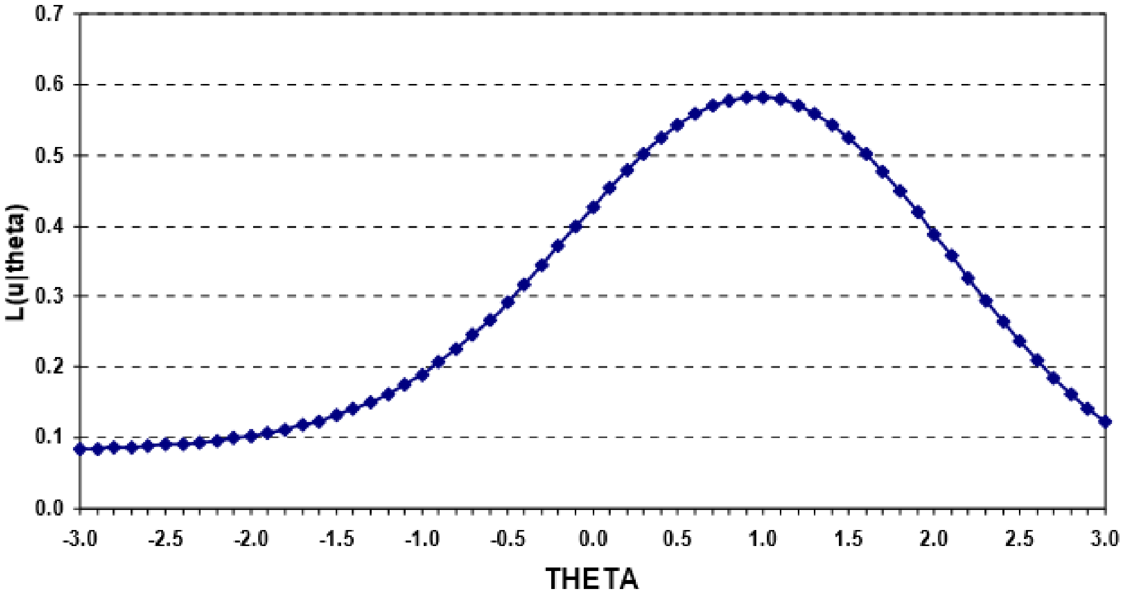
(2)

$$L\left(u \middle| \Theta_j, \beta_1, \dots, \beta_n\right) = \prod_i P_i(\Theta_j)^{u_i} Q_i(\Theta_j)^{1-u_i}$$

In this equation, Q_i represents the probability of an incorrect response to item i and equals $1 - P_i$.

To illustrate the product rule written in Equation 2, we have multiplied the probability of the correct response to Item 1 and the probability of the incorrect response to Item 2 at various trait levels and plotted these values in Figure 5. The result, called the "likelihood function," is single-peaked, with a maximum at Θ = 1.0. This value is called the maximum likelihood estimate of the examinee's trait level (it is the mode of the likelihood function). In other words, it is the most likely trait level given the observed response pattern for the two items. Of course, other estimates are also possible, but Figure 5 shows that the 1.0 value is most likely.

Figure 5: Likelihood Plot.



Finding the Θ value that maximises the conditional probability equation is straightforward. A brute force solution is to calculate the likelihood value at many Θ points across the trait range and simply select the trait level corresponding to the highest likelihood value. This is exactly what was done in the example above. An analytic solution based on an algorithm first suggested by Sir Isaac Newton, which requires fewer calculations, can also be implemented.

A critical problem with maximum likelihood estimates is that no trait level can be estimated for examinees with all-correct or all-incorrect response patterns: the ability estimate is plus or minus infinity. In addition, maximum likelihood ability estimation works best when a relatively large number of items is available. Both of these limitations make the maximum likelihood estimation impractical in a CAT context. Consequently, other trait estimation methods, such as maximum a posteriori (MAP) or expected a posteriori (EAP), are preferred in practice.

EAP estimation (Bock & Mislevy, 1982) is a Bayesian estimator derived from finding the mean of the posterior distribution of the trait given the item responses. The posterior distribution is computed as the conditional probability of the response pattern (see Equation 2) multiplied by the prior distribution function (usually the standard normal function). The prior distribution is simply the probability distribution from which an examinee is drawn. Incorporating this prior into the likelihood function improves trait estimation in cases when all-correct/all-incorrect response patterns are observed or when tests are relatively short. The main advantage of EAP estimation is that it is a non-iterative procedure, so no search methods, such as the brute force or Newton-Raphson method described above, need to be implemented. The equation for EAP estimation is as follows:

(3)

$$\Theta_{EAP} = {\sum_{r=1}^{80} L(Q_r) * Q_r * W(Q_r)} \Big/ {\sum_{r=1}^{80} L(Q_r) * W(Q_r)}$$

Q_r represents the Θ values of 80 equally spaced quadrature nodes (ranging from -4 to 4), the $W(Q_r)$ are the weights at each node taken from a standard normal distribution and $L(Q_r)$ are the conditional probability values of the response pattern at each value of Θ (these are computed using Equation 2). Although it is possible to use more or less than 80 summation nodes in Equation 3, we have found 80 to provide very accurate results and, with modern computers, to be computed very quickly. The resulting trait level estimate represents the mean of the posterior distribution. This EAP estimation method is implemented in AdaptGRT.

As discussed above, an important feature of IRT is that a Θ estimate can be obtained given that there are observations from any set of items with known item parameters. Most significantly, if item parameters for a given item pool are on the same metric, then any subset of items administered to a specific examinee would yield trait estimates on the same scale. In other words, the trait in IRT is item-invariant (of course, trait estimates are more accurate when an appropriately difficult set of items are administered to an examinee). The item invariance property of IRT is critical for CAT applications because examinees do not take the same sequence or number of items.

2.3.5 Evaluating the Accuracy of IRT Trait Estimates

Another important question for CAT applications is: How do the items administered to an examinee affect the precision of the Θ estimate? Although the estimation methods described above result in an estimate of the examinee's trait level, a quick glance at the likelihood plot in Figure 5 suggests that there is a substantial likelihood of other values (the likelihood of nearby trait levels is not much lower than the likelihood at the maximum). The width of the likelihood function near its mode in Figure 5 (i.e., the likelihood function is relatively flat near its maximum) shows that we should not have too much confidence in the obtained maximum likelihood estimate. Intuitively, a likelihood function more sharply peaked would be preferred, because a greatly reduced range of the alternative estimates is possible. The curvature of the likelihood function at its maximum can be used to characterize the precision of the trait estimate. It can be approximated by the test information function (TIF):

(4)

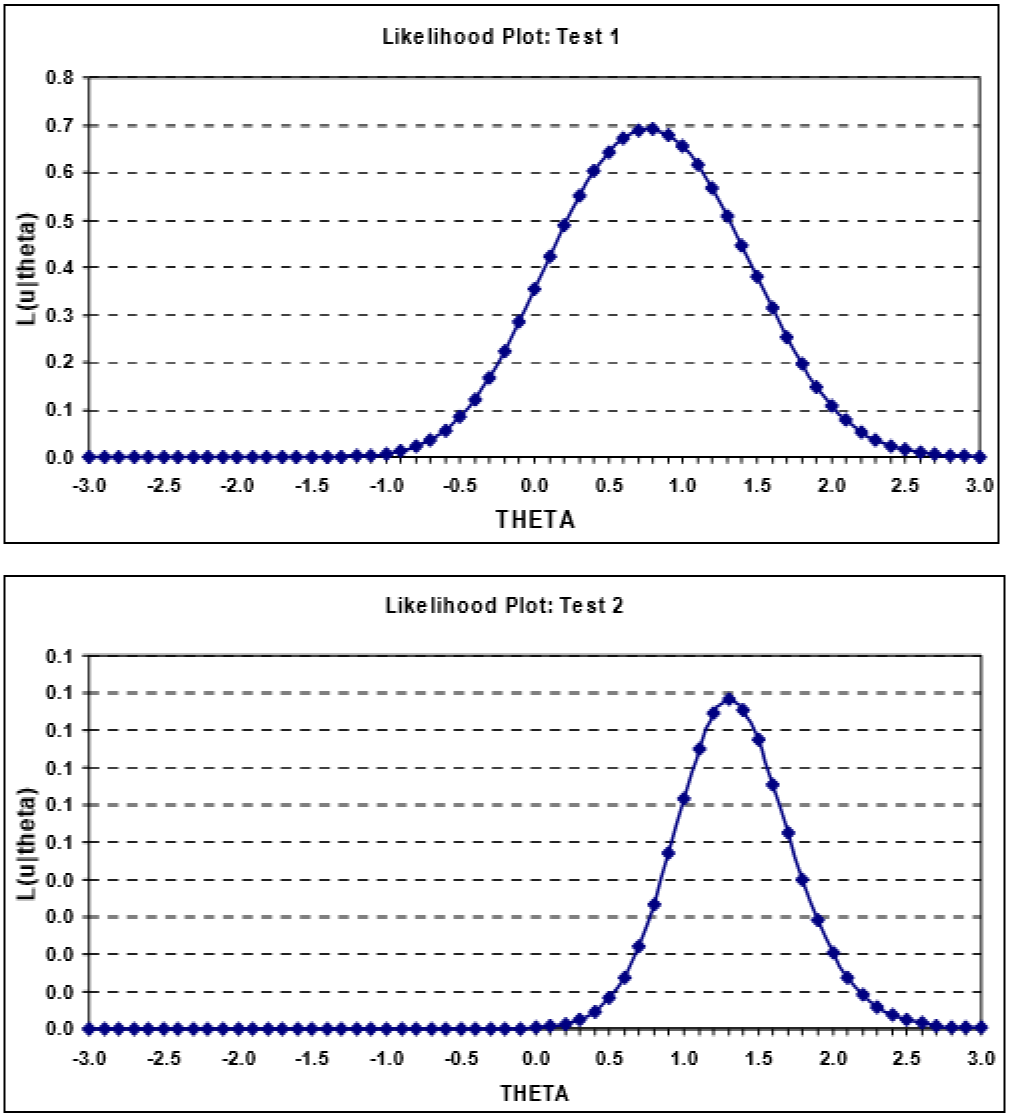
$$I(\Theta) = \sum_{i=1}^{n} \frac{[P_i^{'}(\Theta)]^2}{P_i(\Theta) \cdot Q_i(\Theta)}$$

Whereby $P_i^{'}$ is the first derivative of P_i for item i with respect to Θ. The quantity inside the sum is known as an item information function (IIF) and thus adding the IIFs produces the TIF. The higher the TIF, the narrower the likelihood function. Further inspection of Equation 4 reveals that it is: 1) additive (adding more items produces higher information); and 2) does not depend on item responses, so it can be calculated for the set of items administered to an examinee.

To make the likelihood function sharply peaked (or information high) for a particular examinee, one can increase the number of items administered. However adding items that have flat IRFs at the examinee's trait level does little to increase the curvature of the likelihood function (essentially it multiplies the likelihood by a constant, which will not increase curvature). A more useful strategy is to present items having difficulty parameters similar to that examinee's trait level. Here the slope of the IRF is high and multiplying it into the likelihood increases curvature. CAT applications achieve their efficiency by thoughtfully selecting such items so that the length of the test can be kept to a minimum. To illustrate, Figure 6 presents two hypothetical likelihood functions for an examinee with Θ=1.2 taking two 5-item tests. The first test contained items with a broad range of item difficulties, while the second test contained items having difficulties clustered around that examinee's trait level.

As expected, the width of the likelihood function of the second test is less than the width of the first. In fact, administering a 10 or even 15 item test, similar to test 1 (i.e., having items with a broad range of item difficulties), would not achieve the precision of the second test, even though it is considerably shorter.

Figure 6: Likelihood functions for two 5-item tests.



To summarise, the curvature of the likelihood function tell us about the precision of an ability estimate: greater curvature means greater precision. The TIF tells us how much curvature of the likelihood function is expected at the maximum given the set of items contained in a test. Consequently, the TIF can be computed before a test is administered to assess its precision at various Θ values. The TIF is the sum of the IIFs for the items contained in the test; the IIFs combine additively to form the TIF. Therefore, in a CAT the values of the IIFs at the current ability estimate for the as-yet un-administered items can be computed. The item with the largest IIF can be selected and administered next in the CAT; this is called maximum information item selection.

In the AdaptGRT CAT algorithm, item information functions (IIFs) are computed for each item and stored in "lookup tables." These values are used during CAT administration to determine which items to administer to a particular examinee.

Although the reciprocal of the test information function given in Equation 4 can be used to approximate the error variance of the EAP estimate, in practice, information values are used exclusively for item selection. The standard error of the EAP estimate, called the posterior standard deviation (PSD), can be computed directly by the following equation:

(5)

$$PSD = \sqrt{\frac{\sum_{r=1}^{80}(Q_r - \Theta_{EAP)}{}^{2} * L(Q_r) * W(Q_r)}{\sum_{r=1}^{80} L(Q_r) * W(Q_r)}}$$

This equation is implemented in the AdaptGRT program to evaluate the accuracy of trait estimates for each examinee. A more detailed discussion of adaptive item selection procedure in the AdaptGRT is provided below.

2.3.6 Computerized Adaptive Testing based on the 3PLM

Unlike traditional testing environments where one or more test forms are constructed in advance and items are administered to examinees in a prescribed order, computerized adaptive tests can be constructed on the fly so that each examinee receives a unique set of items that provide near maximum information, in a psychometric sense, at an examinee's estimated trait score at any point during an exam.

As discussed above, test items are chosen to provide near maximum information at the examinee's level of ability. Once each item has been answered, the examinee's ability estimate is updated and provides the basis for the selection of the following item.

Adaptive testing in this fashion is psychometrically efficient, often yielding precision similar to conventional (non-adaptive) tests having nearly twice as many items. In addition, CATs tend to provide higher accuracy and precision at extreme trait levels than non-adaptive tests, which improves their utility for decision making and diagnostic feedback.

To illustrate how items are selected in AdaptGRT in more detail, consider, for example, the item information equation for the 3PLM, shown below:

(6)

$$I_i(\Theta) = \left[a_i^2 \frac{1 - P_i(\Theta)}{P_i(\Theta)}\right] * \left[\frac{(P_i(\Theta) - c_i)^2}{(1 - c_i)^2}\right]$$

Before each item is selected, Equation 6 is used to compute the amount of information provided by each available item in the pool at the examinee’s estimated trait score (EAP). Selecting the item that provides the most information would be optimal in a psychometric sense, but that would lead to items with the largest a-parameters being overused. Thus, in the AdaptGRT, item exposure is controlled by identifying a small subset of items that provide near maximum information at the EAP trait score plus or minus a small random number, and an item is then selected randomly from that subset for presentation to the examinee. Simulation studies comparing non-adaptive and adaptive item selection using this and similar approaches have consistently shown sizeable improvements in testing efficiency. Typically, CAT produces trait scores with similar accuracy and precision using only half as many items.

3. DEVELOPMENT OF THE ADAPTGRT

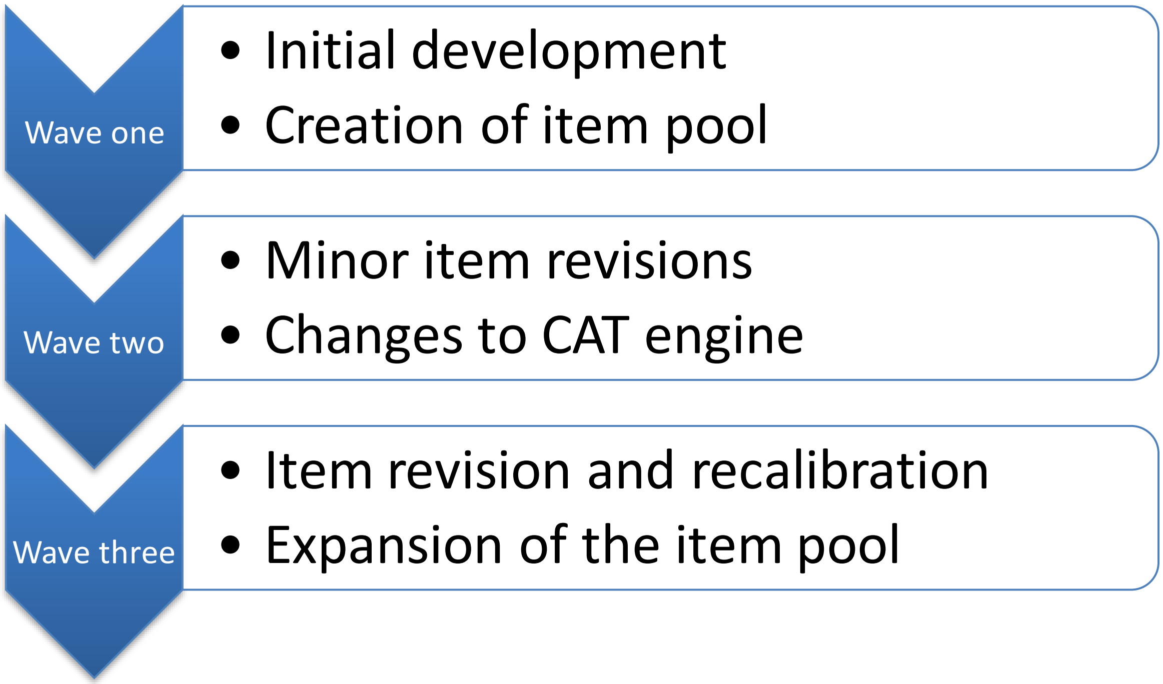
The AdaptGRT was initially adapted from a paper-and-pencil cognitive ability assessment. Since then, it has undergone a number of successive development stages, each of which has aimed to improve the items and functioning of the assessment. This section provides more information into the research methodology and analysis that was undertaken at each stage.

Thus far, there have been three waves of development of the AdaptGRT assessment. The process started in 2010 and continues to date.

The first wave centred around the initial creation of the computer adaptive test, and entailed item creation, pretesting, calibration and the establishment of a robust CAT engine. Wave two focused on addressing minor issues that were identified in the functioning of the test (e.g. item over-exposure) by revising certain items and building additional complexity into the CAT engine.

Wave three entails an on-going development process that aims to refine the existing AdaptGRT items and continually add new items into the test. This cyclical design process and focus on ever-increasing quality is further described in section 3.4 below.

Figure 7: Visual overview of the AdaptGRT development process



3.1 WAVE ONE: INITIAL DEVELOPMENT OF THE ADAPTGRT

The AdaptGRT was originally conceptualised as a computerized adaptive version of Psytech's paper-and-pencil General Reasoning Test Battery or GRT2. The GRT2 was selected as it is a broad-range ability test that is based on multiple years of stringent validation and research (Psytech International Limited, 2010). Both the GRT2 and the AdaptGRT assess three domains that relate to reasoning, namely Verbal, Numerical and Abstract. Each of these tests contain a variety of different item formats to ensure that they provide a broad measure of the dimensions underlying each of the three domains.

Verbal and Numerical ability assess, as their respective names would suggest, the ability to use words and numbers in a rational way, correctly identifying logical relationships between these entities and drawing conclusions and inferences from them. Abstract reasoning assesses the ability to identify logical relationships between abstract spatial relationships and geometric patterns.

3.1.1 Creating and pretesting the item pool

Items for the AdaptGRT were developed and pretested over a period of two years. For each test, a large number of items were written with a similar item stem length. These items had to vary in difficulty in order to provide a measure for examinees at each point of the ability continuum. The items were also written with two primary IRT assumptions in mind, namely unidimensionality and local independence (Kolen & Brennan, 2014).

- Unidimensionality requires that an item or test be related to a single underlying construct.
- Local independence requires that an item on the test is independent of another item. Therefore an examinee's response on item A should not influence his response on item B.

If these assumptions are not taken into account when designing items, they can have a detrimental effect on model fit and parameter estimates. Violation of local independence may lead to an overestimation of reliability and an underestimation of the standard error of estimates.

Because AdaptGRT is predominantly utilised for employee selection and screening applications where testing time is at a premium, efforts were made to keep the content of items as simple and short as possible. Moreover, the types of questions being asked were also kept to the minimum and, wherever possible, similar question types were embedded into multiple tests.

Specifically, all three subsections of the assessment contained the "Associations" question type (e.g., "3.18 is to 1.06 as 1.23 is to …?", "Zinc is to Orchid as Metal is to …?") and the "Odd one out" question type (e.g., "Which of the following is the odd one out :Vehicle, Motorbike, Lorry, Aeroplane, Hovercraft, or Train?"); the Abstract Reasoning and Numerical Reasoning tests contained the "Complete the Sequence" question type (e.g., "1, 3, 5, 6, 10, 9, 15, 12, …What number comes next?"). The Numerical Reasoning test also contained a unique "Problem" question type (e.g., "If New York is five hours behind London and Los Angeles is three hours behind New York, what time is it in Los Angles when it is 4pm in London?"), while the Verbal Reasoning test contained a unique "Antonyms" question type (e.g., "Cry is the opposite of …?"). Regardless of the type, all questions had six (6) response options to minimize the effects of guessing. All these steps were taken to decrease time and effort expended by examinees to familiarize themselves with the testing environment and to standardise test administration.

Once items were written, they were grouped into 20- to 35- item test forms and loaded onto the Psytech assessment platform. Data were predominantly collected via practice tests. Thus, the samples used to pretest AdaptGRT items closely mirrored the population of future examinees. Examinees were randomly assigned to one of the test forms so that the resulting datasets could be treated as equivalent. Additional checks were performed by seeding items in multiple test forms and comparing item p-values across datasets.

Although no demographic information was collected during pretesting due to the confidentiality agreement (i.e., examinees were given an informed consent prior to the practice test informing them that the data would be used only for test development purposes), as a group, the majority of examinees were applicants for skilled-level positions in the United Kingdom, with about equal percentage of males and females, and a diverse mix of cultural and racial backgrounds.

Table 1 presents item pool and sample size statistics for each of the three AdaptGRT tests. As can been seen, a total of 268 Verbal Reasoning items were pretested by forming 11 test forms. Note that because of the item overlap, a total of 306 items were actually pretested. For the Numerical Reasoning test, 177 unique items were pretested in 10 test forms, while for the Abstract Reasoning test, there 126 unique items in 5 test forms. Although the sample sizes varied between 200 and 5000 across pretest forms, the majority of samples exceeded 500 examinees, which is sufficient for accurate IRT parameter estimation provided that marginal maximum likelihood or the corresponding Bayesian estimation method is used. Note that even when sample sizes were relatively small (i.e., IRTN33 test form), the data were merged with a much larger dataset (i.e., IRTN3) and concurrent item parameter calibration was performed.

3.1.2 Item Coding and Pre-calibration

Classical Test Theory (CTT) item analyses were conducted to verify the quality and appropriateness of the items for the three-parameter logistic (3PL) IRT model (see chapter two). Specifically, p-values and corrected item total correlations (CITCs) for each item were computed and examined to determine whether they were correctly coded and had positive correlations with other items in the test form. Items with low p-values (i.e., when the percent correct was at or below the chance level of .15) were checked for possible content or coding problems and, if necessary, deleted from the pool. Items with negative or near zero CITCs were also deleted; such items were either misinterpreted by examinees or contained more than one correct answer and would likely cause program convergence problems during IRT item calibration.

As a result of pre-calibration item analyses 11 Numerical Reasoning items were identified as problematic and deleted from the pool. Altogether, however, the initial Numerical item pool performed extremely well and less than 7 % of initial items (11 out of 177) were screened out prior to IRT analyses. Items for the Abstract Reasoning test also performed well. Only 6 out of 126 initial items were identified as problematic. Finally, for the Verbal Reasoning, 26 items were identified as problematic representing 10% of the initial pool of 266. All problem items were deleted, yielding the 240 item subset to be calibrated.

Table 1. Pretest item pool and sample size statistics for the three AdaptGRT tests.

Pool Name	# Of Items	Sample Size	Notes
		Verbal Ability	
IRTV1	25	2231	Merged with GRTV1 because of item overlap
IRTV2	25	421	Merged with GRTV2 because of item overlap
IRTV3	22	786	Merged with GRTV3 because of item overlap
IRTV4	35	5170	
IRTV5	30	2506	
IRTV11	26	563	
IRTV22	26	1087	
IRTV33	27	733	
GRT1VR	30	1286	Merged with IRTV1 because of item overlap
GRT2VR	30	1581	Merged with IRTV2 because of item overlap
GRT3VR	30	3325	Merged with IRTV3 because of item overlap
Total Items	306		
Total Unique Items	268		
		Numerical Ability	
IRTN1	24	576	Merged with IRT11 and GRTN1 because of item overlap
IRTN2	22	843	Merged with IRT22 and GRTN2 because of item overlap
IRTN3	20	6739	Merged with IRT33 because of item overlap
IRTN4	25	2962	
IRTN5	25	1774	
IRTN11	30	564	All items are from IRTN1 or GRTN1 subsets
IRTN22	30	205	All items are from IRTN2 or GRTN2 subsets
IRTN33	30	241	Merged with IRTN3 because of item overlap
GRTN1	30	879	Merged with IRTN1 and IRNT11 because of item overlap
GRTN2	30	1229	Merged with IRTN2 and IRTN22 because of item overlap
Total Items	266		

Total Unique Items	177		
Abstract Ability			
IRTA1	30	3064	Shares 1 item with IRTA3
IRTA2	26	1385	
IRTA3	21	362	Shares 1 item with IRTA1
IRTA4	25	753	
IRTA5	25	1271	
Total Items	127		
Total Unique Items	126		

3.1.3 Item Calibration

The items of the AdaptGRT were calibrated according to the 3PL IRT model, using the BILOG-MG for Windows computer programme (duToit, 2003). In an ideal world, the entire item pool would have been completed by the same heterogeneous sample group so that parameters could be simultaneously calibrated for all items. However, given the large number of items, this was not feasible in practice and therefore the item pool was divided into smaller subsets. These subsets shared common items and therefore it was possible to merge the data at a later stage, coding missing responses as "not reached" and concurrently calibrating item parameters. In situations where subsets did not share common items, the pretest sample was assumed to be equivalent in ability distribution, due to being randomly assigned to subset groups and the total sample size often exceeding 1000 examinees.

In total, the research team conducted seven separate item parameter calibrations for the Verbal Reasoning test, 5 calibrations for the Numerical Reasoning test, and five calibrations for the Abstract Reasoning test. Data that were missing by design were treated as "not reached." Default priors were used for the discrimination and difficulty parameters, numerical integration was performed with 40 quadrature points, and the convergence criterion was set to 0.01 as recommended by BILOG. All calibration runs converged before maximum number of E-M cycles (100) was reached.

Model-data fit was evaluated using both graphical (fit plots) and statistical (chi-squares for item singles, doubles, and triples) methods (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). As expected, the 3PL fit AdaptGRT items very well. With just few exceptions, adjusted chi-square to degrees of freedom ratios for item singles, doubles and triples were below or at the 3.0 level recommended by Drasgow, Levine, Tsien, Williams, and Mead (1995). In rare cases when fit statistics exceeded the recommended cut-off levels, misfitting items were deleted.

The resulting 3PL item parameters were next examined to evaluate their suitability for CAT. Items with low discrimination parameters (below 0.40) were deleted. These items tend to provide little information at any trait level and are unlikely to be selected by the adaptive testing algorithm. Items with moderate discrimination parameters (between .40 and .50) were only retained if their difficulty parameters were in the middle to positive trait ranges. In these regions, there were relatively few items available, so these items would stand a higher chance of being used during the CAT.

The final item set for Verbal Reasoning consisted of 200 items, while Numerical Reasoning retained a total of 155 items and Abstract Reasoning 110 items. The summary statistics for each item set are presented in Table 2 below. As can be seen, each test contained a diverse set of items with varying discrimination, difficulty, and guessing parameters. The Numerical and Abstract Reasoning tests had larger proportions of items with lower difficulty parameters than the Verbal Reasoning test, but there were sufficient numbers of items in the middle and high trait ranges in all tests.

Table 2. Summary of IRT and CTT statistics for the three AdaptGRT item sets.

AdaptGRT Test	No. of Items		3PLM Parameters			CTT Statistics	
			a	b	c	CITC	p-value
Verbal Test	200	min	0.42	-4.80	0.01	0.13	15.00
		max	3.26	4.07	0.50	0.58	98.90
		mean	0.80	-0.37	0.15	0.35	76.75
		sd	0.33	1.32	0.07	0.10	20.55
Numerical Test	155	min	0.42	-3.89	0.05	0.13	4.30
		max	2.41	2.53	0.30	0.58	98.80
		mean	0.94	-1.22	0.16	0.35	62.81
		sd	0.33	1.35	0.04	0.10	21.58
Abstract Test	110	min	0.42	-3.24	0.03	0.12	15.60
		max	1.59	3.09	0.46	0.56	97.20
		mean	0.88	-0.75	0.17	0.36	71.86
		sd	0.26	1.07	0.06	0.09	18.00

3.1.4 Test Information Functions (TIF) and Standard Error (SE)

As discussed in chapter two, the item information function (IIF) and corresponding aggregated TIF provide an indication of the 'information' that the item or test provides when predicting an examinee's ability. According to Baker (2008), information can be considered an IRT analogue of classical test theory's (CTT's) reliability measure and provides an indication of how precisely the test can distinguish between those with higher or lower levels of ability.

The Standard Error is another indicator of item quality and can be considered an analogue of the standard error of measurement typically calculated in CTT analyses. As the information of a test increases, the SE of the test should decrease. The same is true for the information function and SE of items. In an ideal test, items would determine the maximal amount of information about an examinee's ability with the smallest amount of SE. Information functions are also useful when demonstrating which areas of the trait continuum individual items and the item pool as a whole are measuring well and in which areas they are not.

These functions were plotted for the three AdaptGRT tests once calibration was complete. As can be seen from figure 8, Verbal Reasoning items possessed higher information curves and lower standard errors across a wider range of trait values than the other two tests. This was partly due to the fact that the item pool for the Verbal Reasoning test contained 200 items as opposed to 155 items in the Numerical Reasoning item pool and 110 items in the Abstract Reasoning item pool. Additionally, item difficulties were distributed more evenly for this item pool so there were enough highly discriminating items for all trait regions. For Numerical and Abstract Reasoning, the TIF indicated that the items discriminated best at the 1.5 to 2.0 trait region – see figures 9 and 10. Both of these tests contained a small number of difficult items, causing the standard error to increase at the high end of the ability range. However, because the primary purpose of the AdaptGRT is to screen applicants at low and middle trait ranges, all three item pools were judged to be sufficiently heterogeneous for CAT purposes.

Figure 8. Test Information Function (TIF) and Standard Error (SE) of the Verbal Reasoning test.

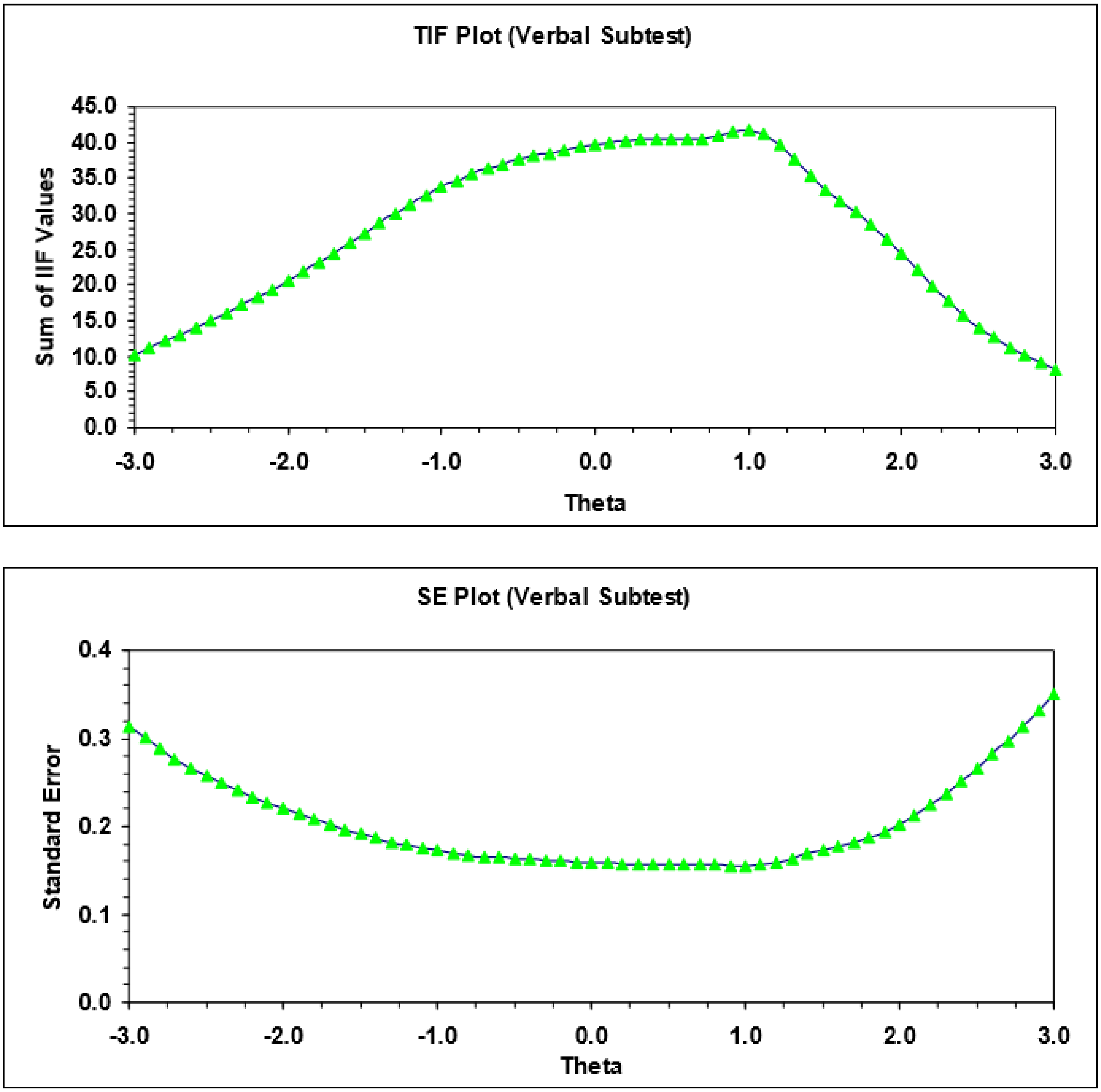


Figure 9. Test Information Function (TIF) and Standard Error (SE) of the Numerical Reasoning test.

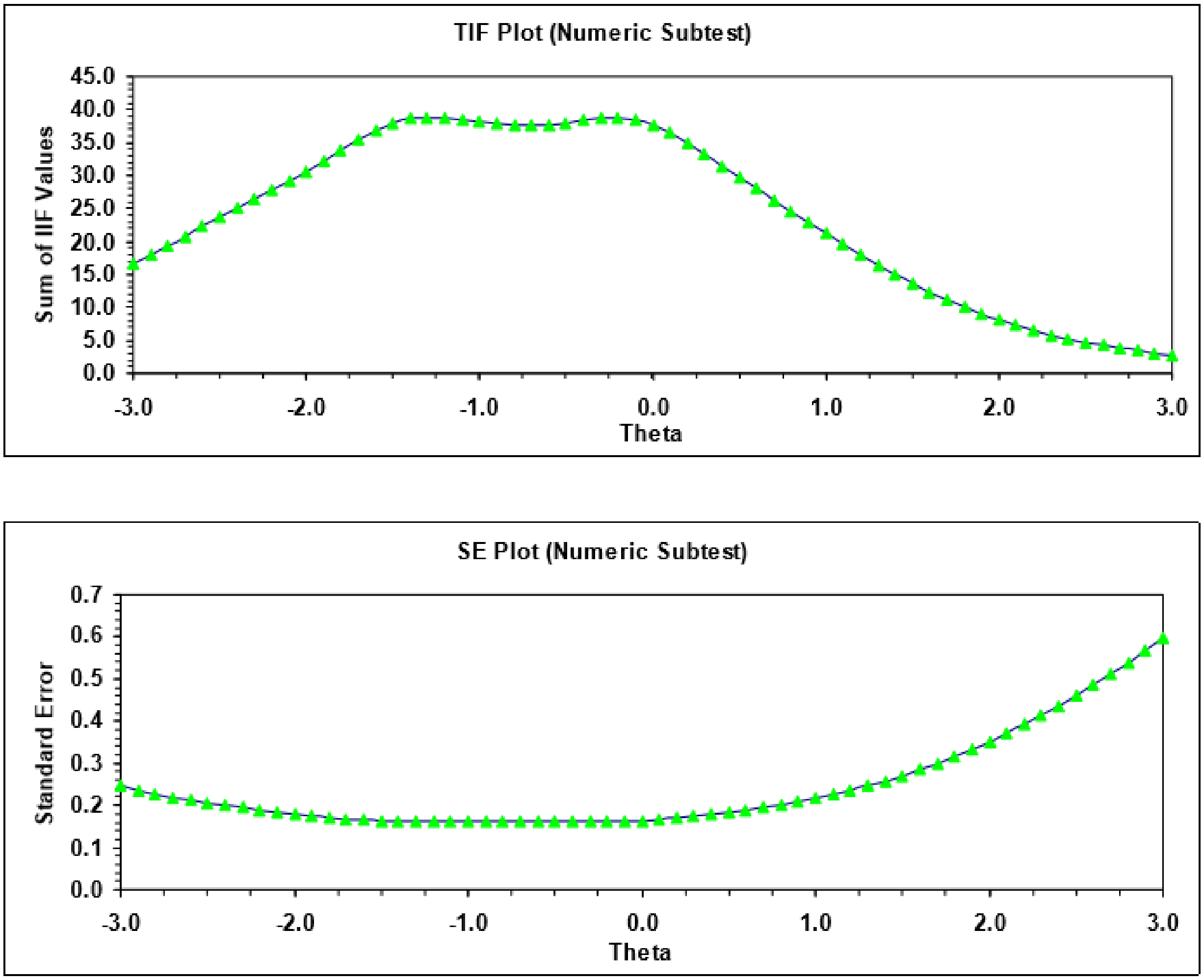
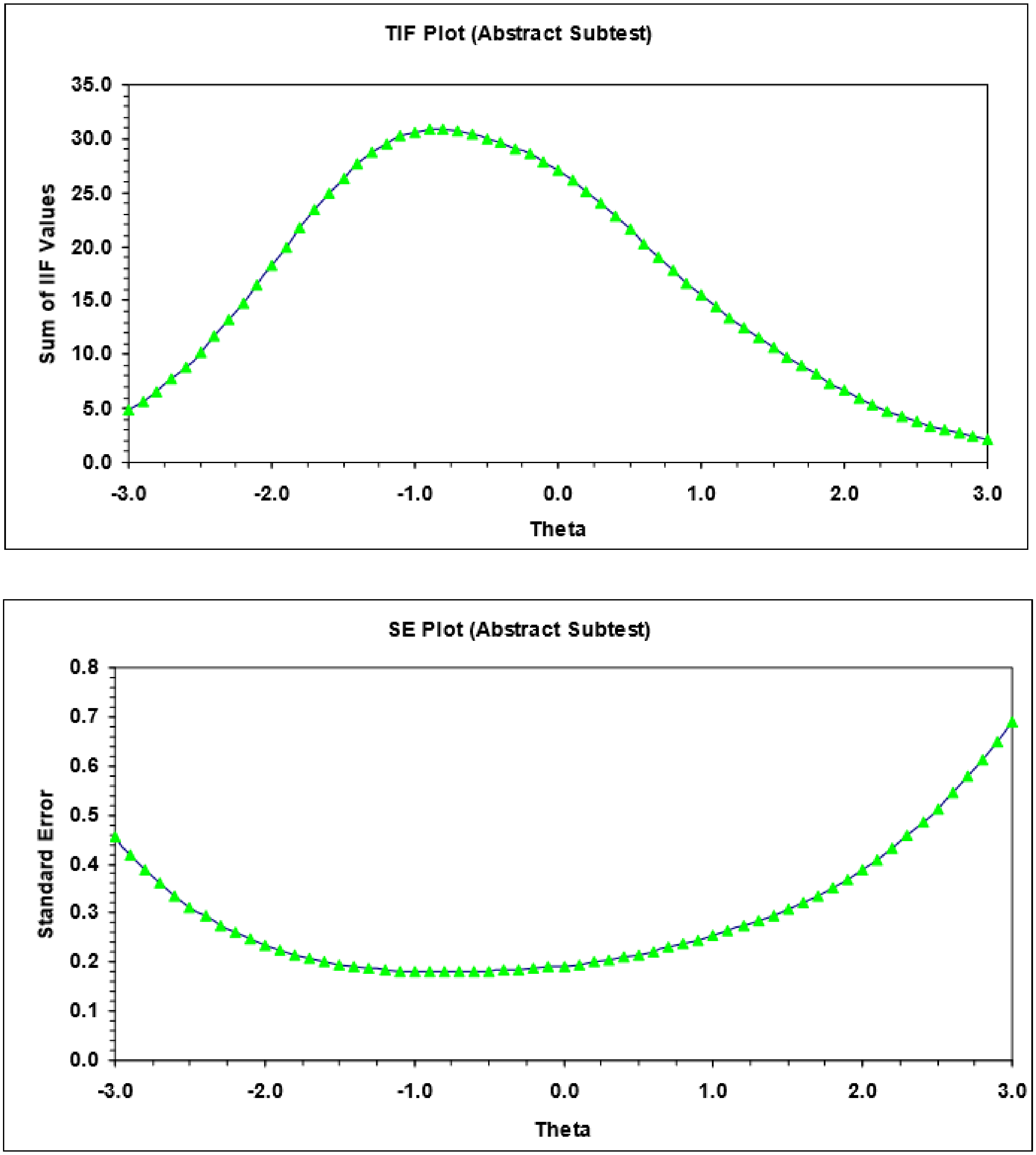


Figure 10. Test Information Function (TIF) and Standard Error (SE) of the Abstract Reasoning test.



3.1.5 Trait Estimation and Norming

One of the key advantages of IRT over CTT methods is that its trait estimates are item invariant. This means that, as long as test items are placed on the same metric, IRT trait score estimates are directly comparable regardless of which subset was actually administered to an examinee. As discussed in chapter two, the AdaptGRT used EAP as a method of trait estimation.

Item responses were captured and merged to create a single dataset for each test. Responses were transformed into a dichotomous variable with 0 denoting an incorrect response and 1 a correct response. 9 was used to represent those items that were not administered. Each examinee had between 20 and 35 responses, which were sufficient for accurate scoring. Next, these responses, along with the corresponding IRT item parameter estimates, were submitted into Stark's 3PL_EAP computer program, scored, and saved into an Excel spreadsheet for further analyses.

It was then decided to normalise the EAP estimates. This is not necessary within the IRT paradigm but becomes valuable when reporting on results in practice, as standardised scales are more readily understood and interpreted within a business setting. Normalising also allows for the computation of a composite score that provides an indication of overall test performance.

To create norms for each AdaptGRT test, distribution statistics and plots based on the EAP scores were computed. A conversion table was then created in order to determine percentile ranks, stanine and Z-scores (see Table 3). EAP score frequency distributions for the three tests are shown in Figure 11 below.

As can be seen, the frequency distributions closely approximated the normal distribution throughout much of the range of scores. However, the tails of the distributions clearly diverged from the standard normal distribution. The conversion table depicted in table 3 allows EAP scores to be "normalised" by looking up the ability estimate, finding the percentile score, and using a "p-to-z" transformation that gives the standard normal z-score of the corresponding percentile value. Thereafter the z-score can be transformed to any desired score reporting scale. For example, if T-scores are desired with a mean of 50 and a standard deviation of 10, the simple transformation T = 10z + 50 can be used. The overall General Reasoning score is a unit weighted composite of the z-scores or T-scores from the three AdaptGRT subtests.

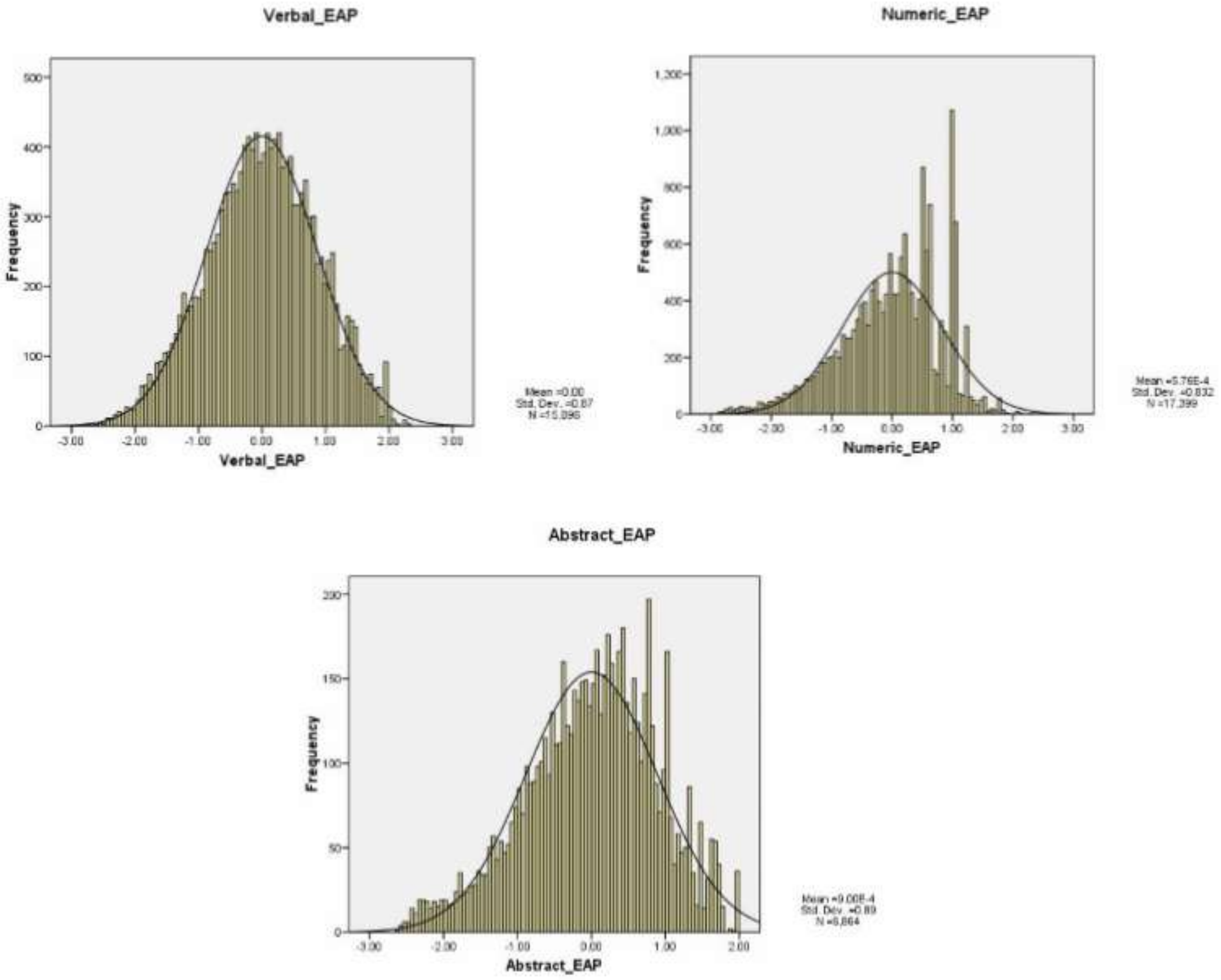
Table 3. EAP score conversion table for AdaptGRT tests.

	IRT EAP Scores				Reporting		
	Numerical	**Abstract**	**Verbal**	**Stanines**	**Percentiles**	**z-score**	**T-score**
N	17399	6864	15096				
Mean	0.00	0.00	0.00				
SD	0.83	0.89	0.87				
Variance	0.69	0.79	0.76				
Skewness	-0.59	-0.35	-0.12				
Kurtosis	0.14	-0.17	-0.32				
Minimum	-2.86	-2.61	-2.75				
Maximum	2.31	1.96	2.30				
	-2.321	-2.275	-2.018	1	1	-2.33	26.74
	-1.997	-2.087	-1.818		2	-2.05	29.46
	-1.793	-1.877	-1.683		3	-1.88	31.19
	-1.658	-1.755	-1.589		4	-1.75	32.49
	-1.541	-1.618	-1.491	2	5	-1.64	33.55
	-1.435	-1.501	-1.414		6	-1.55	34.45
	-1.347	-1.400	-1.340		7	-1.48	35.24
	-1.270	-1.332	-1.276		8	-1.41	35.95
	-1.196	-1.264	-1.226		9	-1.34	36.59
	-1.138	-1.195	-1.179		10	-1.28	37.18
	-1.082	-1.129	-1.124		11	-1.23	37.73
	-1.028	-1.064	-1.071	3	12	-1.17	38.25
	-0.976	-1.026	-1.023		13	-1.13	38.74
	-0.928	-0.979	-0.976		14	-1.08	39.20
	-0.881	-0.936	-0.926		15	-1.04	39.64
	-0.832	-0.891	-0.887		16	-0.99	40.06
	-0.796	-0.855	-0.849		17	-0.95	40.46
	-0.758	-0.815	-0.812		18	-0.92	40.85
	-0.718	-0.781	-0.777		19	-0.88	41.22
	-0.680	-0.743	-0.742		20	-0.84	41.58
	-0.641	-0.706	-0.708		21	-0.81	41.94
	-0.605	-0.668	-0.675		22	-0.77	42.28
	-0.575	-0.638	-0.648		23	-0.74	42.61
	-0.541	-0.608	-0.617	4	24	-0.71	42.94
	-0.513	-0.576	-0.586		25	-0.67	43.26
	-0.488	-0.540	-0.557		26	-0.64	43.57
	-0.462	-0.516	-0.534		27	-0.61	43.87
	-0.436	-0.485	-0.506		28	-0.58	44.17
	-0.402	-0.454	-0.478		29	-0.55	44.47
	-0.373	-0.425	-0.452		30	-0.52	44.76
	-0.340	-0.395	-0.426		31	-0.50	45.04
	-0.317	-0.374	-0.396		32	-0.47	45.32
	-0.295	-0.352	-0.372		33	-0.44	45.60
	-0.271	-0.324	-0.347		34	-0.41	45.88
	-0.254	-0.296	-0.321		35	-0.39	46.15
	-0.228	-0.265	-0.298		36	-0.36	46.42

	-0.204	-0.241	-0.276	5	37	-0.33	46.68
	-0.175	-0.219	-0.253		38	-0.31	46.95
	-0.150	-0.191	-0.230		39	-0.28	47.21
	-0.119	-0.166	-0.208		40	-0.25	47.47
	-0.092	-0.142	-0.187		41	-0.23	47.72
	-0.069	-0.118	-0.164		42	-0.20	47.98
	-0.051	-0.096	-0.141		43	-0.18	48.24
	-0.033	-0.070	-0.118		44	-0.15	48.49
	-0.014	-0.050	-0.096		45	-0.13	48.74
	0.006	-0.022	-0.074		46	-0.10	49.00
	0.035	0.001	-0.054		47	-0.08	49.25
	0.058	0.023	-0.026		48	-0.05	49.50
	0.081	0.046	-0.005		49	-0.03	49.75
	0.101	0.069	0.017		50	0.00	50.00
	0.137	0.088	0.043		51	0.03	50.25
	0.160	0.112	0.066		52	0.05	50.50
	0.165	0.139	0.088		53	0.08	50.75
	0.190	0.164	0.107		54	0.10	51.00
	0.208	0.185	0.129		55	0.13	51.26
	0.239	0.207	0.150		56	0.15	51.51
	0.239	0.225	0.174		57	0.18	51.76
	0.254	0.247	0.198		58	0.20	52.02
	0.290	0.268	0.222		59	0.23	52.28
	0.299	0.288	0.242		60	0.25	52.53
	0.329	0.311	0.267	6	61	0.28	52.79
	0.341	0.338	0.286		62	0.31	53.05
	0.388	0.354	0.306		63	0.33	53.32
	0.415	0.382	0.332		64	0.36	53.58
	0.427	0.401	0.356		65	0.39	53.85
	0.464	0.416	0.381		66	0.41	54.12
	0.488	0.436	0.405		67	0.44	54.40
	0.506	0.460	0.428		68	0.47	54.68
	0.506	0.488	0.452		69	0.50	54.96
	0.508	0.509	0.477		70	0.52	55.24
	0.515	0.544	0.502		71	0.55	55.53
	0.547	0.570	0.528		72	0.58	55.83
	0.565	0.590	0.556		73	0.61	56.13
	0.572	0.612	0.588		74	0.64	56.43
	0.599	0.642	0.620		75	0.67	56.74
	0.615	0.679	0.644		76	0.71	57.06
	0.655	0.706	0.669		77	0.74	57.39
	0.655	0.732	0.693	7	78	0.77	57.72
	0.655	0.753	0.720		79	0.81	58.06
	0.706	0.777	0.749		80	0.84	58.42
	0.781	0.788	0.783		81	0.88	58.78
	0.800	0.804	0.814		82	0.92	59.15
	0.849	0.836	0.843		83	0.95	59.54
	0.867	0.874	0.885		84	0.99	59.94

	0.950	0.915	0.918		85	1.04	60.36
	0.988	0.957	0.958		86	1.08	60.80
	0.988	0.999	0.999		87	1.13	61.26
	0.988	1.013	1.045		88	1.17	61.75
	0.988	1.035	1.081		89	1.23	62.27
	0.988	1.075	1.122		90	1.28	62.82
	0.995	1.129	1.155		91	1.34	63.41
	1.025	1.197	1.221		92	1.41	64.05
	1.025	1.274	1.291	8	93	1.48	64.76
	1.025	1.303	1.360		94	1.55	65.55
	1.043	1.371	1.423		95	1.64	66.45
	1.203	1.460	1.485		96	1.75	67.51
	1.237	1.590	1.580		97	1.88	68.81
	1.303	1.650	1.711	9	98	2.05	70.54
	1.517	1.740	1.900		99	2.33	73.26

Figure 11. Frequency distributions for AdaptGRT EAP scores.



3.1.6 Test-retest Reliability

A CTT analysis into the reliability of the AdaptGRT was also conducted to determine whether ability estimates remained comparable across time. A study conducted within a New Zealand governmental agency found evidence of good to excellent test-retest reliability (Brown & McInnes, 2001). For each of the three AdaptGRT tests, scores across testing sessions closely resembled one another. The correspondence between the test and retest scores on the AdaptGRT's overall composite was outstanding.

The sample size consisted of 85 adult participants, of which 47 completed the AdaptGRT twice in a period of between five to nine weeks. The remaining 38 participants failed to complete the retest for multiple reasons. After data cleaning removed those answers that showed evidence of inattentiveness, there were 38 remaining cases for the Abstract and Numerical tests and 45 remaining cases for the Verbal test.

Factor analyses were performed on the data to assess whether the overall composite test score could accurately be used in place of the individual test scores. The conclusion was that a single factor could explain most of the variance in test scores, and that this factor consisted of approximately equal loadings of Abstract, Numerical, and Verbal scores. Consequently, it is justifiable to use a single composite score that is a simple sum of the Abstract, Numerical, and Verbal Reasoning scores.

Test-rest analysis was conducted on z-scores that were based on the New Zealand norm group with a mean of 0 and a standard deviation (SD) of 1. Pearson's correlation coefficient was used to assess the correspondence between the two sets of scores, on an individual test level as well as on the composite score. The minimum acceptable correlation between test and retest scores is 0.7 (see Kline, 2000). A correlation coefficient of 0.8 is considered good. Table 4 shows the correlations found between all test and retest scores

Table 4: Correlation coefficients for AdaptGRT scores during test and retest.

	Retest			
Test	**Abstract**	**Numerical**	**Verbal**	**Composite**
Abstract	**0.76**	0.53	0.58	0.74
Numerical	0.71	**0.85**	0.52	0.81
Verbal	0.59	0.64	**0.84**	0.80
Composite	0.82	0.79	0.77	**0.93**

Note: correlations in bold are test-retest correlations.

Table 4**Error! Reference source not found.** shows that the test-retest correlation coefficients for the Abstract, Numerical, and Verbal tests were *r* = 0.76, 0.85, and 0.84, respectively. In other words, test-retest reliability was fair for the abstract reasoning subtest, and very good for the numerical and verbal reasoning subtests. These test-retest reliability coefficients are similar or slightly superior to published coefficients (Psytech International, 2010) for a non-adaptive reasoning test, the GRT2. Test-retest reliability for the composite score was outstanding: $r = 0.93$.

The researchers further found that the scores of the second test sitting were not subject to the typical memory effects seen in non-adaptive tests. This was commended as an advantage of the AdaptGRT, in that the content of the test changes at each sitting.

3.2 WAVE TWO: MINOR REVISIONS AND CHANGES TO THE CAT ENGINE

Once the research process undertaken during wave one was complete, the final parameter estimations were added to the CAT engine and the AdaptGRT was put into practice. Over time, certain issues came to the fore that sparked the need for a second development phase. These issues included:

- **Coverage of the ability range:** The majority of the items on the three AdaptGRT tests were focused around the lower to average ability range. There was a need to broaden the item pool to include more difficult items in order to assess those at the higher end of the trait continuum.
- **Item overexposure:** Certain items with steep IIFs were overexposed during the testing session, especially those within the average trait range. These were continually routed by the item selection algorithm as, statistically, they provided the most information into an examinee's ability level.
- **Time limit:** During the pretesting phase of wave one, examinees completed a static subset of items within a practice test environment. The time limits of these sessions were applied to the AdaptGRT tests and resulted in multiple examinees being unable to complete the test in the allotted period.
- **Low stakes vs. high stakes:** Item pretesting occurred in a low stakes environment while the AdaptGRT was typically used in high stakes testing. This fundamental change in the nature of the test environment led to examinees typically scoring lower on the AdaptGRT, compared to their scores on the practice tests. It also resulted in the parameters of certain items changing between the pilot study and practical application of the test.

3.2.1 Expansion of the item pool and updated norms

Due to these issues being raised, it was decided to calibrate the items once more using the BILOG-MG programme for Windows. Those items that were highlighted as being problematic were amended to address issues of multidimensionality, confusing item stems or a lack of clarity around the correct answer.

The research team also transformed the EAP scores into a standard normal distribution. This shifted the score axis in order to improve examinee performance and address the issue of test-takers scoring lower than expected. A test taker's EAP score therefore became a reflection of his/her ability relative to the scale of normal distribution. Norm conversion tables (see table 3) were also updated in order to reflect the EAP scores of the revised items and facilitate their conversion into various standardised scales.

3.2.2 Changes to the CAT engine

In order to address the issue of item overexposure, a randomisation parameter was added to the CAT engine. Previously the selection algorithm overexposed items that discriminated very precisely between different levels of ability. The amended algorithm ensured that each presented item was randomly selected from a range of possible items, yet still reported high degrees of information. The CAT engine was also updated to include the revised parameter estimates obtained during the second round of item calibration.

3.3 WAVE THREE: ITEM RECALIBRATION AND REVISION

The changes undertaken during wave two provided a short-term solution to the concerns identified with the AdaptGRT. Transforming the EAP scores allowed participants to receive higher scores but did not alter the initial issue of examinee results being lower than expected. In addition, there still remained a paucity of items in certain areas of the ability range. Therefore, an additional, comprehensive round of test validation was undertaken and is currently under way.

The goal of wave three was to optimise functioning of the AdaptGRT by identifying items that were still problematic after the first two stages of development, in terms of parameters, model fit, information values, SE and multicollinearity.

3.3.1 Initial Item Recalibration

All data relating to previous validations was merged into a single dataset. Examinee responses to the items were recoded from the Likert scale response format to dichotomous variables where 0 represented an incorrect answer and 1 a correct answer. Items were recalibrated using the IRTPRO software package (Cai, Thissen & Du Toit, 2011). Once again, all items were analysed according to the 3PLM (see chapter two).

In order to recalibrate the items, analysis was undertaken per test (Verbal, Numerical or Abstract). Within each test, the total number of validated items was split into subsets, each of which was analysed separately. These subsets were identical to those provided in Table 1 and were derived from the original validation test forms. The Verbal Ability test, with a total of 200 validated items, was split into 11 subsets. The 155 items of the Numerical Ability test were split into 8 subsets, while the 110 Abstract items were divided into 5 subsets. Each subset consisted of between 9 to 28 items.

Findings demonstrated parameter estimates and fit indices that were significantly different to those recorded in the original validation process. Specifically, researchers found differences in the a, b and c parameters within the Verbal, Numerical and Abstract tests, with changes in a and c values markedly more substantial than changes in b values. Items were also found to be problematic for the following reasons:

- **Poor model fit:** Items were assessed in terms of their degree of fit to the 3PLM. This was determined using a chi-squared statistic, with significance ($p>0.05$) or values above 20 indicating poor model fit. 67% of previously validated items were found to have a chi-squared index above

20. Certain items within the Abstract Reasoning test were found to be markedly problematic with chi-squared values higher than 100.

- **Low information values:** Multiple items were found to have low information values with relatively flat IIFs. Such findings imply that the items do not contribute a significant amount of information in the estimation of examinee ability and are not of great value within a CAT assessment.
- **High SE:** Along with flat information curves, certain items reported high SE values on a parameter level.
- **Multicollinearity:** Items that demonstrated multicollinearity were found to be highly correlated and, as such, contributed relatively little unique variance to the assessment.

3.3.1.1 Verbal ability

Table 5 provides a summary of the mean parameter changes for each subset of Verbal ability items.

Table 5: Summary of Revised Parameters and Model Fit for Verbal Ability.

Subset name	No. of Items	Fit	Mean a			Mean b			Mean c		
			Original	Revised	Change	Original	Revised	Change	Original	Revised	Change
IRTV1	18	60.32	0.70	1.19	0.48	-0.10	1.09	1.19	0.14	0.19	0.04
GRT1VR	13	20.19	0.76	1.25	0.49	0.66	-0.27	-0.93	0.17	-0.17	-0.33
IRTV2	21	36.91	1.13	1.93	0.79	0.08	-0.68	-0.75	0.16	-0.84	-1.00
GRT2VR	20	46.86	0.71	1.16	0.45	0.06	-0.17	-0.22	0.15	0.24	0.09
IRTV3	18	11.68	0.76	1.32	0.56	-1.07	-1.03	0.04	0.13	1.60	1.47
GRT3VR	9	111.28	0.79	2.10	1.30	-0.71	1.38	2.09	0.14	1.09	0.96
IRTV4	27	179.85	0.73	1.16	0.42	-0.85	-0.35	0.50	0.16	0.93	0.77
IRTV5	21	64.23	0.83	1.25	0.42	-0.04	-0.26	-0.22	0.16	0.24	0.07
IRTV11	14	22.02	0.86	1.42	0.56	-0.28	0.10	0.38	0.14	0.39	0.26
IRTV22	16	55.97	0.74	1.03	0.29	-0.02	-0.78	-0.77	0.16	0.05	-0.11
IRTV33	22	34.20	0.81	1.95	1.14	-1.37	-0.66	0.71	0.16	1.36	1.20
MEAN TOTAL	**199**	**58.50**	**0.80**	**1.43**	**0.63**	**-0.33**	**-0.15**	**0.18**	**0.15**	**0.46**	**0.31**

The overall average fit parameter for the Verbal Ability items was 58.50, a value somewhat higher than the cut-off point of 20. Two subsets contained items that were considerably problematic in terms of fit, namely GRT3VR (111.28) and IRTV4 (179.85). In an effort to improve the quality of the Verbal Ability subscale, 104 out of a total of 199 items were revised or suggested for deletion.

The a parameter increased by an average of 0.63, implying that the items had steeper discrimination curves than originally proposed.

Changes to the b parameter were minimal with an increase of 0.18. Overall, Verbal Ability items were found to be relatively easy to answer with a mean total difficulty parameter value of -0.15. The b parameter of the IRTV1 subset changed dramatically from -0.10 to 1.09; this set of items was found to be more difficult than previously indicated. The GRT3VR subset also saw an increase in the difficulty of items.

Item recalibration for the c parameter led to minimal changes with an overall increase of 0.31. C-parameters were a cause for concern with many items being susceptible to guessing or manipulation. Mean c parameters were especially high on the following subsets: IRTV33 (1.36), IRTV3 (1.60) and GRT3VR (1.09).

Based on the results for the Verbal test, a large proportion of the items required revision with the goal of broadening the range of the item pool to provide greater coverage at the higher end of the trait continuum.

3.3.1.2 Numerical Ability

Table 6 provides a mean summary of the parameter estimation changes in the item subsets of Numerical Ability.

Table 6: Summary of Revised Parameters and Model Fit for Numerical Ability.

Sub-set name	No. of items	Fit		Mean a			Mean b			Mean c	
			Original	Revised	Change	Original	Revised	Change	Original	Revised	Change
IRTN3	15	26.11	0.93	1.66	0.73	-2.25	-2.12	0.13	0.14	3.28	3.15
IRTN33	17	12.09	1.20	2.11	0.91	-0.64	-0.42	0.22	0.18	1.04	0.86
IRTN2	17	38.43	0.84	1.36	0.52	-1.18	-2.17	-0.98	0.17	1.57	1.40
GRT2NR	17	24.46	1.02	1.50	0.48	0.44	0.02	-0.43	0.13	-0.35	-0.48
IRTN1	21	15.95	0.89	1.88	0.99	-2.34	-1.63	0.71	0.17	3.00	2.82
GRT1NR	14	108.17	1.05	1.73	0.68	-1.09	-0.97	0.12	0.17	1.62	1.45
IRTN4	19	33.52	0.84	1.74	0.90	-1.72	-1.25	0.47	0.15	2.12	1.98
IRTN5	23	31.67	0.91	1.65	0.75	-0.88	-0.31	0.57	0.16	1.14	0.98
MEAN TOTAL	**143**	**36.30**	**0.96**	**1.70**	**0.75**	**-1.21**	**-1.11**	**0.10**	**0.16**	**1.68**	**1.52**

The overall parameter changes for the Numerical Ability items were minimal. Overall fit was also more optimal than that seen in the Verbal Ability test with an overall chi-squared value of 36.30. Out of a total of 143 items, 68 were earmarked for revision.

On average, the α parameter increased by 0.75, implying that the items demonstrated greater power to discriminate between individuals with different levels of ability. This suggests that the slopes of the sinusoidal curves are much steeper than originally anticipated, with certain items discriminating more precisely along a narrower trait range.

There were minimal changes in the b parameter between the item calibration undertaken during wave one and wave three. With a mean b parameter of -1.11, the items for Numerical Ability were situated on the easier end of the trait continuum. On average the b values were negative in most subsets with the exception of the GRT2NR subset. This may suggest gaps in the item pool on the higher end of the trait continuum.

Similarly to the Verbal Ability items, the c parameters for the second stage of calibration changed by 1.52 units. With an overall mean c value of 1.68, items demonstrated vulnerability to guessing or manipulation by examinees across all levels of ability.

3.3.1.3 Abstract Ability

The mean parameter values for Abstract Ability, compared across the first and third waves of item calibration, are provided in table 7.

Table 7: Summary of Revised Parameters and Model Fit for Numerical Ability.

Sub-set name	No. of items	Fit		Mean a			Mean a			Mean a	
			Original	Revised	Change	Original	Revised	Change	Original	Revised	Change
IRTA1	28	190.49	0.87	1.68	0.81	-0.64	0.06	0.70	0.14	0.52	0.38
IRTA2	18	80.99	0.78	6.17	5.39	-0.90	0.78	1.69	0.17	-4.41	-4.58
IRTA3	20	29.68	0.94	10.11	9.17	-0.87	1.43	2.30	0.19	-14.20	-14.38
IRTA4	20	153.77	0.88	6.07	5.19	-0.59	0.34	0.93	0.19	-1.86	-2.05
IRTA5	24	70.25	0.94	6.45	5.51	-0.85	0.88	1.73	0.17	-5.40	-5.57
MEAN TOTAL	**110**	**105.04**	**0.88**	**6.10**	**5.21**	**-0.77**	**0.70**	**1.47**	**0.17**	**-5.07**	**-5.24**

Overall, the items on Abstract Ability were found to be significantly more problematic than previously demonstrated. The model fit statistic, in particular, was found to be far higher than expected with an overall mean value of 105.04. The IRTA3 subset, with a chi-squared value of 29.68 was found to be the best functioning overall. The mean a value seen in this analysis (6.10) was problematic and may have been influenced by the disproportionate model fit parameters.

The b parameters changed from being negative during the first round of calibration (-0.77) to positive (0.70). These changes were not as substantial as the changes seen in the a and c parameters.

The c value also changed sizably, with a mean revised parameter value of -5.07. Negative c values are problematic in nature, suggesting items that are open to manipulation and guessing responses.

3.3.2 Item Deletion and Revision

At this point items that were found to be problematic were deleted or revised. In certain instances issues were traced to vague item stems or response options without a clear answer or more than one correct answer. In other cases, the items contained terms that were not culturally neutral and may have been unfamiliar to certain individuals. These terms were substituted for words that are more universally applicable.

For the Verbal test, 52% of the existing item pool required revision, based on the results of the item recalibration, while 48% of the Numerical items were revised. The amount of Abstract items undergoing revision has yet to be determined.

3.3.2.1 Item revision methodology

Once the problematic items were isolated, a panel of subject-matter experts (SMEs) were consulted in order to analyse the existing item content and amend the items accordingly. Where suitable, the items were kept as similar as possible to their existing content. Of particular importance was culture fairness – ensuring that the item stem and answer options contained content that is recognisable to participants from diverse backgrounds.

The revision process was conducted as follows:

- The two lead researchers captured the original item and suggested a revised item stem and/or answer options.
- The original items and suggested revisions were sent onto the SME panel.
- The panel independently reviewed the suggestions for objectivity, language and culture fairness.
- Only those revised items that received a consensus from the reviewers were retained; the other items were deleted.

3.3.2.2 Development of additional items

Beyond the list of items to be revised, there was also a requirement to develop new, additional items for the Verbal, Numerical and Abstract Ability subtests. As there were significantly fewer Abstract items to be reviewed, compared to Verbal and Numerical Ability, this was pinpointed as the first area of priority for the development of new items. This was done in order to expand on the number of item formats available within the Abstract test, as well as to balance the number of Verbal, Numerical and Abstract items earmarked for revision. The items were initially created by the two lead researchers and reviewed by the panel in a similar process to the one discussed in section 3.3.2.1.

3.3.3 Item seeding

Once the revision process was complete, the revised and/or new items (Abstract only) were grouped together into sub-sets of 7 items. All items within a set of 7 had the same question format (e.g. *what number comes next?).*

3.3.3.1 Item anchors and parameter drift

Each set of 7 items included some trialling items, as well as two to four anchor items. Anchor items are validated, well-performing items with established parameter estimates calibrated to a particular measurement scale. These are included in order to provide a reference point when estimating the parameters of the trial items (Wise & Kingsbury, 2000).

The inclusion of anchor items is done in order to prevent the occurrence of item parameter drift (IPD). According to Babcock & Albano (2012), IPD describes fluctuations in item parameter estimates that may occur across different groups of test takers. Unchecked, IPD may have a serious impact on the measurement precision of trait estimates obtained from the computer adaptive test.

In some instances it was not possible to include anchor items into the item set. This was due to a lack of existing items with the same underlying item format, e.g. when new items and a new item format was developed.

3.3.3.2 Trialling methodology

In order to obtain accurate and robust parameter estimates, it was important to trial the items within a high stakes environment. Often, a pilot study on new or revised items is conducted in static, paper-and-pencil, low-stakes conditions, yielding statistical outputs that are not reflective of the conditions in which the test will be used.

IIn order to address this need, the revised items were seeded into an existing, validated computer-based cognitive assessment, known as the Internet Reasoning Test (IRT3). The IRT3 test, also created and sold by Psytech International, follows the same structure as the Adapt GRT assessment with three subtests of reasoning ability: Verbal, Numerical and Abstract. Many of the IRT3 questions also follow the same underlying item formats as the AdaptGRT questions. It is important to note that while the AdaptGRT revised items were trialled within high-stakes testing conditions, they were not trialled in a CAT environment.

3.3.3.3 Item delivery and scoring

The original trialling methodology allowed participants who were completing the full IRT3 assessment (with 17 items per subtest) to also complete an additional 7 trial items. These were seeded into the online cognitive test and blended into the overall testing process. While participants completed a total of 24 items per subtest, only the validated IRT3 items were scored and presented in the report.

Unfortunately the original process was judged to be too time consuming and a second approach was then proposed and accepted. This process reduced the number of validated IRT3 items presented to participants to 10 out of a total of 17 per subtest. An analysis was completed demonstrating that this number of items could still provide a robust, statistically sound result for each of three subtests. In addition to these 10 validated items, an additional 14 trial items were presented to each participant within the testing environment. Participants still completed a total of 24 items per subtest, but were only scored on the 10 validated items per subtest.

An additional strategy to speed up data collection was to trial all revised items at the same time. The sets of 7 revised items were reorganised into pools of 14 items, typically consisting of 7 items of one format and 7 items of a different format – see Table 8 for a visual overview. Participants who completed the IRT3 were automatically assigned one of the item pools for each subtest.

Table 8. Overview of item pools for Verbal, Numerical and Abstract Ability.

Item pool	No. of trial items	No. of IRT3 items
IT1 (original methodology)	7	17
IT2	14	10
IT3	14	10
IT4	14	10
IT5	14	10
IT6	14	10

3.3.3.4 Item analysis

As with the analysis discussed in section 3.3.1, IRTPRO version 3 was used to analyse the trial items and estimate the parameters according to the 3PLM model. Datasets were organised per subtest (Verbal, Numerical and Abstract) and per item pool (the group of 14 trial items that was randomly assigned to a particular group of participants). The IRT3 items completed by the participants were also included in each dataset.

When estimating the item parameters and fit, separate analyses were conducted for each item format present within a particular item pool. This meant that there were typically two separate analyses per dataset. Where the IRT3 questions also corresponded to the relevant item format, they were also included into the analysis with the view to including them into the AdaptGRT assessment.

With reference to section 3.3.3.1, the a parameter for the established anchor items was constrained to the existing value in each analysis. This was done in order to prevent the occurrence of IPD by ensuring that all parameters were estimated on the same measurement scale.

3.3.4 Revised item parameter estimates

In each of the subsections below, the parameters are provided for each of the item pools and associated item formats.

3.3.4.1 Verbal Ability revised estimates

Tables 9 to 17 depict the a, b and c parameters, as well as goodness of fit, for the anchor and trial items within each item pool (IT1 to IT6) and item format. The parameters and fit of the previously problematic items are compared to the updated revised items to track changes.

The anchor items within each table are coloured in grey. Where the ^ symbol occurs, the updated fit of the item could not be calculated due to limited degrees of freedom.

IT1 (n=233):

Table 9. V_IT1 Odd one out - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
V1R1_1A	IRTV1_09	0,81	0,81	0,00	1,62	-1,65	-3,27	0,17	0,18	0,01	22,43	11,15	-11,28
V1R1G_3A	GRT1VR_4	1,01	1,01	0,00	-0,35	0,73	1,08	0,12	0,09	-0,03	14,57	14,15	-0,42
V1R1_4	IRTV1_03	0,87	0,86	-0,01	-0,52	0,39	0,91	0,08	0,15	0,07	34,39	9,94	-24,45
V1R1_5	IRTV1_05	1,02	1,17	0,15	-0,34	1,28	1,62	0,28	0,16	-0,12	40,22	16,65	-23,57
V1R1_6	IRTV1_06	0,52	0,62	0,10	-1,06	-1,28	-0,22	0,18	0,19	0,01	62,22	9,00	-53,22
V1R1_7	IRTV1_08	0,66	1,15	0,49	-2,14	-2,01	0,13	0,14	0,19	0,05	300,61	5,52	-295,09
IRT3V1_1			2,20			-1,15			0,20			5,06	
IRT3V1_2			0,97			-0,30			0,18			12,28	
IRT3V1_3			0,91			-1,25			0,18			7,89	
IRT3V1_4			1,49			-0,29			0,18			6,79	
IRT3V1_5			1,87			-0,21			0,17			9,68	
IRT3V1_6			1,85			1,61			0,10			10,51	
IRT3V1_7			1,49			0,95			0,13			13,34	

IT2 (n=324):

Table 10. V_IT2 Odd one out - 3PL Model Item Parameter Estimates.

Trial	Original	Original	Revised	Change	Original	Revised	Change	Original	Revised	Change	Original	Revised	Change
Name	name	a	a	a	b	b	b	c	c	c	fit	fit	fit
V2R1_1A	IRTV1_09	0,81	0,81	0,00	1,62	-1,46	-3,08	0,17	0,13	-0,04	22,43	22,42	-0,01
V2R1G_3A	GRT1VR_04	1,01	1,01	0,00	-0,35	-0,25	0,10	0,12	0,09	-0,03	14,57	26,44	11,87
V2R1G_4	GRT1VR_10	0,48	1,24	0,76	0,58	-0,33	-0,91	0,26	0,17	-0,09	35,27	8,51	-26,76
V2R1G_5	GRT1VR_24	0,53	0,28	-0,25	-1,32	-9,12	-7,80	0,15	0,20	0,05	45,46	4,92	-40,54
V2R1G_6	GRT1VR_30	0,81	0,71	-0,10	-0,39	0,36	0,75	0,18	0,18	0,00	23,87	14,80	-9,07
V2R1G_7	GRT1VR_18	0,52	1,78	1,26	0,66	0,39	-0,27	0,09	0,16	0,07	13,26	5,70	-7,56
IRT3V2_1			3,31			-1,05			0,17			5,21	
IRT3V2_2			1,57			-1,27			0,19			6,28	
IRT3V2_5			2,01			-0,26			0,22			8,37	
IRT3V2_7			1,48			-1,25			0,18			10,37	
IRT3V2_10			1,70			-0,77			0,24			12,00	

Table 11. V_IT2 XYZ- 3PL Model Item Parameter Estimates.

Trial	Original	Original	Revised	Change	Original	Revised	Change	Original	Revised	Change	Original	Revised	Change
Name	name	a	a	a	b	b	b	c	c	c	fit	fit	fit
V3R1_1A	IRTV2_12	0,51	0,51	0,00	-0,42	0,64	1,06	0,10	0,13	0,03	14,27	18,49	4,22
V3R1_2A	IRTV2_13	1,41	1,41	0,00	0,72	1,66	0,94	0,10	0,06	-0,04	16,01	6,51	-9,50
V3R1G_3A	GRT2VR_14	0,67	0,67	0,00	1,69	2,00	0,31	0,10	0,09	-0,01	17,80	12,65	-5,15
V3R1_4	IRTV2_01	0,98	1,09	0,11	0,44	1,53	1,09	0,16	0,15	-0,01	65,71	2,33	-63,38
V3R1_5	IRTV2_03	0,67	1,16	0,49	-2,06	-2,11	-0,05	0,18	0,20	0,02	39,60	6,26	-33,34
V3R1_6	IRTV2_05	2,52	0,18	-2,34	1,82	-7,52	-9,34	0,01	0,20	0,19	102,96	9,17	-93,79
V3R1_7	IRTV2_07	0,53	2,02	1,49	-1,56	-0,84	0,72	0,18	0,20	0,02	72,22	5,75	-66,47
IRT3V2_4			1,73			-1,63			0,20			3,66	
IRT3V2_9			1,66			-0,14			0,19			1,71	

IT3 (n=315):

Table 12. V_IT3 What comes next - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
V5R1_1A	IRTV3_09	0,67	0,67	0,00	-2,35	-2,12	0,23	0,14	0,18	0,04	15,85	16,66	0,81
V5R1_2A	IRTV3_14	0,59	0,59	0,00	-0,54	0,10	0,64	0,08	0,10	0,02	7,49	20,16	12,67
V5R1_3A	IRTV3_21	0,57	0,57	0,00	1,11	1,15	0,04	0,26	0,21	-0,05	4,31	7,54	3,23
V5R1G_4	GRT3VR_09	0,86	1,74	0,88	-0,56	-0,25	0,31	0,15	0,16	0,01	77,36	11,46	-65,90
V5R1G_5	GRT3VR_26	0,96	2,23	1,27	-0,92	-0,37	0,55	0,07	0,14	0,07	46,04	13,94	-32,10
V5R1_6	IRTV33_04	0,59	3,13	2,54	-1,42	-0,89	0,53	0,18	0,19	0,01	33,70	9,26	-24,44
V5R1_7	IRTV33_10	0,59	1,55	0,96	-1,87	-0,84	1,03	0,17	0,19	0,02	29,86	9,69	-20,17
IRT3V3_3			1,48			-1,02			0,16			10,62	
IRT3V3_6			2,12			-1,21			0,18			7,92	
IRT3V3_8			2,05			0,21			0,13			11,17	

Table 13. V_IT3 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
V4R1_1A	IRTV2_12	0,51	0,51	0,00	-0,42	0,47	0,89	0,10	0,10	0,00	14,27	25,21	10,94
V4R1_2A	IRTV2_13	1,41	1,41	0,00	0,72	1,16	0,44	0,10	0,07	-0,03	16,01	5,89	-10,12
V4R1G_3A	GRT2VR_14	0,67	0,67	0,00	1,69	1,68	-0,01	0,10	0,09	-0,01	17,80	13,98	-3,82
V4R1_4	IRTV2_20	0,56	1,30	0,74	-0,66	-2,14	-1,48	0,21	0,20	-0,01	45,53	8,22	-37,31
V4R1G_5	GRT2VR_12	0,72	2,30	1,58	-1,39	-0,75	0,64	0,14	0,19	0,05	41,96	4,73	-37,23
V4R1G_6	GRT2VR_16	0,89	1,37	0,48	0,56	0,93	0,37	0,08	0,15	0,07	44,14	3,45	-40,69
V4R1G_7	GRT2VR_19	0,52	0,67	0,15	0,57	-0,34	-0,91	0,28	0,20	-0,08	76,07	3,79	-72,28
IRT3V3_4			0,89			-0,69			0,19			9,67	
IRT3V3_9			2,95			0,04			0,15			4,21	

IT4 (n=311):

Table 14. V_IT4 Odd one out - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
V6R1_1A	IRTV1_09	0,81	0,81	0,00	1,62	-1,49	-3,11	0,17	0,16	-0,01	22,43	19,32	-3,11
V6R1G_3A	GRT1VR_04	1,01	1,01	0,00	-0,35	-0,09	0,26	0,12	0,10	-0,02	14,57	14,44	-0,13
V6R1_4	IRTV4_13	0,91	1,74	0,83	0,70	0,58	-0,12	0,07	0,14	0,07	120,19	13,41	-106,78
V6R1_5	IRTV4_29	0,50	1,02	0,52	-1,22	-1,42	-0,20	0,10	0,21	0,11	55,80	6,58	-49,22
V6R1_6	IRTV4_35	1,15	0,79	-0,36	0,91	1,55	0,64	0,12	0,14	0,02	51,58	6,55	-45,03
V6R1_7	IRTV5_10	1,08	1,17	0,09	1,24	1,33	0,09	0,09	0,13	0,04	50,79	9,28	-41,51
IRT3V1_1			3,53			-1,27			0,17			6,63	
IRT3V1_2			1,61			-0,51			0,21			9,56	
IRT3V1_5			1,09			-1,84			0,19			11,36	
IRT3V1_7			1,46			-1,04			0,21			7,75	
IRT3V1_10			2,84			-0,42			0,18			4,94	

Table 15. V_IT4 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
V7R1_1A	IRTV2_12	0,51	0,51	0,00	-0,42	0,61	1,03	0,10	0,14	0,04	14,27	7,66	-6,61
V7R1_2A	IRTV2_13	1,41	1,41	0,00	0,72	1,25	0,53	0,10	0,05	-0,05	16,01	8,46	-7,55
V7R1G_3A	GRT2VR_14	0,67	0,67	0,00	1,69	2,03	0,34	0,10	0,09	-0,01	17,80	5,97	-11,83
V7R1_4	IRTV4_14	0,56	1,10	0,54	-4,80	-5,37	-0,57	0,16	0,20	0,04	217,31	^	^
V7R1_5	IRTV4_18	0,86	2,91	2,05	0,00	-0,19	-0,19	0,05	0,14	0,09	29,23	3,10	-26,13
V7R1_6	IRTV5_08	0,53	1,05	0,52	-2,06	-1,00	1,06	0,16	0,19	0,03	210,24	9,16	-201,08
V7R1_7	IRTV5_19	0,74	0,55	-0,19	2,02	-0,46	-2,48	0,50	0,19	-0,31	91,76	3,60	-88,16
IRT3V1_4			1,61			-0,55			0,16			4,62	
IRT3V1_9			1,96			-0,96			0,19			2,07	

IT5 (n=297):

Table 16. V_IT5 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
V8R1_1A	IRTV2_12	0,51	0,51	0,00	-0,42	0,08	0,50	0,10	0,11	0,01	14,27	22,43	8,16
V8R1_2A	IRTV2_13	1,41	1,41	0,00	0,72	1,16	0,44	0,10	0,06	-0,04	16,01	14,59	-1,42
V8R1G_3A	GRT2VR_14	0,67	0,67	0,00	1,69	1,76	0,07	0,10	0,10	0,00	17,80	10,94	-6,86
V8R1_4	IRTV5_22	0,54	0,57	0,03	-1,56	-2,95	-1,39	0,15	0,20	0,05	61,11	11,25	-49,86
V8R1_5	IRTV22_07	0,71	1,97	1,26	-2,41	-1,46	0,95	0,16	0,21	0,05	188,62	8,27	-180,35
V8R1_6	IRTV22_15	0,51	1,63	1,12	-0,52	-0,89	-0,37	0,17	0,18	0,01	35,89	15,43	-20,46
V8R1_7	IRTV22_19	1,08	2,20	1,12	-2,05	-2,19	-0,14	0,18	0,19	0,01	233,59	3,83	-229,76
V12R1_4	IRTV2_11	0,54	0,60	0,06	-0,66	-0,65	0,01	0,17	0,21	0,04	47,09	7,00	-40,09
V12R1_5	IRTV2_18	3,26	1,03	-2,23	1,05	-1,61	-2,66	0,23	0,19	-0,04	25,23	11,00	-14,23
V12R1_6	GRT2VR_27	0,78	1,14	0,36	0,79	-0,27	-1,06	0,07	0,17	0,10	28,07	11,91	-16,16
V12R1G_7	GRT2VR_22	0,58	0,75	0,17	-0,83	0,08	0,91	0,12	0,18	0,06	46,40	13,20	-33,20
V13R1_4	IRTV2_24	0,82	0,62	-0,20	0,30	1,81	1,51	0,14	0,17	0,03	21,63	12,64	-8,99
V13R1_5	GRT2VR_01	0,55	1,13	0,58	1,32	0,99	-0,33	0,15	0,17	0,02	19,83	15,73	-4,10
V13R1_6	IRTV22_14	0,64	1,93	1,29	-0,86	-0,42	0,44	0,18	0,17	-0,01	35,80	7,18	-28,62
IRT3V2_9			1,09			-0,34			0,17			14,16	
IRT3V2_11			1,99			0,27			0,13			10,06	
IRT3V2_14			0,58			0,14			0,19			16,24	
IRT3V2_15			1,25			0,74			0,14			11,47	
IRT3V2_16			2,48			1,17			0,08			9,32	

IT6 (n=307):

Table 17. V_IT6 Odd one out - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
V9R1_1A	IRTV1_09	0,81	0,81	0,00	1,62	-1,32	-2,94	0,17	0,15	-0,02	22,43	13,74	-8,69
V9R1G_3A	GRT1VR_04	1,01	1,01	0,00	-0,35	0,23	0,58	0,12	0,07	-0,05	14,57	23,66	9,09
V9R1_4	IRTV1_17	0,55	0,88	0,33	1,38	-2,89	-4,27	0,17	0,20	0,03	39,66	5,36	-34,30
V9R1_5	IRTV1_18	0,68	2,70	2,02	-1,32	0,34	1,66	0,18	0,09	-0,09	31,89	19,50	-12,39
V9R1_6	IRTV1_07	0,57	1,90	1,33	0,27	1,22	0,95	0,08	0,06	-0,02	21,15	9,46	-11,69
V9R1G_7	IRTV1_11	0,92	2,52	1,60	1,03	-0,03	-1,06	0,24	0,12	-0,12	36,90	15,46	-21,44
V10R1_4	IRTV1_17	0,55	1,82	1,27	1,38	0,99	-0,39	0,17	0,08	-0,09	39,66	21,25	-18,41
V10R1_5	IRTV1_18	0,68	1,01	0,33	-1,32	-3,16	-1,84	0,18	0,20	0,02	31,89	6,57	-25,32
V10R1_6	IRTV1_07	0,57	1,35	0,78	0,27	0,34	0,07	0,08	0,20	0,12	21,15	13,03	-8,12
V10R1_7	IRTV1_11	0,92	0,54	-0,38	1,03	1,37	0,34	0,24	0,20	-0,04	36,90	25,62	-11,28
V11R1G_4	GRT1VR_25	0,95	0,93	-0,02	4,07	-0,94	-5,01	0,07	0,19	0,12	16,02	25,37	9,35
V11R1_5	IRTV5_13	0,71	1,80	1,09	2,13	-0,48	-2,61	0,19	0,18	-0,01	32,50	14,63	-17,87
V11R1_6	IRTV11_11	0,77	1,75	0,98	1,90	-0,26	-2,16	0,11	0,18	0,07	54,06	8,80	-45,26
IRT3V3_10			1,87			-0,04			0,14			9,18	
IRT3V3_12			1,30			0,76			0,11			16,29	
IRT3V3_13			1,18			1,36			0,13			15,33	

As seen in the Verbal Ability trial item tables above, the majority of items demonstrated far better fit, indicating the success of the revisions. The revised a and b values showed good coverage at various levels of the underlying trait.

3.3.4.2 Numerical Ability revised estimates

The revised parameter estimates for the Numerical Ability subtest are depicted in Tables 18 to xx. Each item pool and its associated item formats is depicted separately as follows:

- IT1: Table 18
- IT2: Table 19 and 20
- IT3: Table 21 and 22
- IT4: Table 23 and 24
- IT5: Table 25
- IT6: Table 26 and 27

As above, the anchor items within each table are demarcated in grey. As discussed in section 3.3.3, the a parameter for each of these anchor items was constrained to its CAT engine value, in order to ensure that the updated parameter estimates for the revised items are on the same measurement scale as the existing items.

Where the ^ symbol occurs, the updated fit of the item could not be calculated due to limited degrees of freedom.

IT1 (n=229):

Table 18. N_IT1 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N1R1_1A	IRTN33_21	1,04	1,04	0,00	-2,67	-2,35	0,32	0,20	0,20	0,00	4,85	0,56	-4,29
N1R1_2A	IRTN33_25	1,34	1,34	0,00	1,41	-1,24	-2,65	0,12	0,27	0,15	9,42	22,94	13,52
N1R1_3A	IRTN33_26	1,09	1,09	0,00	0,23	0,96	0,73	0,20	0,11	-0,09	9,91	11,59	1,68
N1R1_4A	IRTN33_28	1,44	1,43	-0,01	-0,53	-0,32	0,21	0,23	0,13	-0,10	9,38	5,11	-4,27
N1R1_5	IRTN3_04	0,91	0,57	-0,34	-3,55	-3,98	-0,43	0,18	0,20	0,02	23,72	^	^
N1R1_6	IRTN3_06	0,93	1,91	0,98	-3,01	-1,13	1,88	0,12	0,19	0,07	49,87	5,42	-44,45
N1R1_7	IRTN3_10	1,03	3,39	2,36	-2,85	-1,44	1,41	0,13	0,19	0,06	48,62	^	^
IRT3N1_3			1,17			-1,40			0,19			5,76	
IRT3N1_7			2,23			-0,78			0,18			4,80	
IRT3N1_11			1,30			-0,82			0,18			2,34	
IRT3N1_16			2,28			-0,26			0,17			2,71	

IT2 (n=301):

Table 19. N_IT2 Odd one out - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N3R1_1A	IRTN2_06	0,57	0,57	0,00	-3,19	-3,25	-0,06	0,18	0,19	0,01	14,87	8,88	-5,99
N3R1_2A	IRTN2_17	1,08	1,08	0,00	1,22	1,61	0,39	0,16	0,08	-0,08	16,57	9,49	-7,08
N3R1_3A	IRTN2_20	1,20	1,20	0,00	0,24	0,70	0,46	0,09	0,08	-0,01	12,51	11,19	-1,32
N3R1G_4A	GRT2NR_17	1,86	1,86	0,00	-0,15	0,31	0,46	0,17	0,13	-0,04	18,68	8,04	-10,64
N3R1_5	IRTN2_12	0,52	0,27	-0,25	0,85	-5,66	-6,51	0,25	0,20	-0,05	33,54	2,88	-30,66
N3R1G_6	GRT2NR_23	0,58	0,40	-0,18	-0,98	-1,27	-0,29	0,11	0,20	0,09	39,20	8,40	-30,80
N3R1G_7	GRT2NR_26	0,42	0,33	-0,09	0,89	-0,96	-1,85	0,19	0,20	0,01	37,09	9,31	-27,78
IRT3N2_5			1,30			-1,36			0,20			5,05	
IRT3N2_10			0,38			1,83			0,20			7,76	

Table 20. N_IT2 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N2R1_1A	IRTN33_21	1,04	1,04	0,00	-2,67	-2,03	0,64	0,20	0,17	-0,03	4,85	14,61	9,76
N2R1_2A	IRTN33_25	1,34	1,34	0,00	1,41	0,57	-0,84	0,12	0,63	0,51	9,42	14,59	5,17
N2R1_3A	IRTN33_26	1,09	1,09	0,00	0,23	0,94	0,71	0,20	0,08	-0,12	9,91	6,78	-3,13
N2R1_4A	IRTN33_28	1,44	1,44	0,00	-0,53	-0,31	0,22	0,23	0,11	-0,12	9,38	7,44	-1,94
N2R1_5	IRTN3_15	0,88	3,60	2,72	-1,51	-0,47	1,04	0,08	0,11	0,03	37,31	7,55	-29,76
N2R1_6	IRTN3_18	0,60	6,84	6,24	-0,06	-0,44	-0,38	0,05	0,08	0,03	-	7,15	-
N2R1_7	IRTN3_20	0,88	0,99	0,11	-2,40	-1,22	1,18	0,14	0,18	0,04	22,25	6,06	-16,19
IRT3N2_3			1,22			-1,32			0,19			4,09	
IRT3N2_7			1,68			-1,45			0,20			1,68	

IT3 (n=301):

Table 21. N_IT3 Odd one out - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N4R1_1A	IRTN2_06	0,57	0,57	0,00	-3,19	-4,37	-1,18	0,18	0,19	0,01	14,87	10,37	-4,50
N4R1_2A	IRTN2_17	1,08	1,08	0,00	1,22	1,76	0,54	0,16	0,11	-0,05	16,57	10,70	-5,87
N4R1_3A	IRTN2_20	1,20	1,20	0,00	0,24	0,84	0,60	0,09	0,09	0,00	12,51	9,74	-2,77
N4R1G_4A	GRT2NR_17	1,86	1,86	0,00	-0,15	0,46	0,61	0,17	0,13	-0,04	18,68	18,30	-0,38
N4R1_5	IRTN2_05	0,88	3,01	2,13	-3,44	-2,23	1,21	0,20	0,19	-0,01	262,48	^	^
N4R1_6	IRTN2_08	0,60	0,98	0,38	-1,85	-1,38	0,47	0,14	0,19	0,05	22,57	11,10	-11,47
N4R1_7	IRTN2_09	0,78	1,13	0,35	-2,98	-1,66	1,32	0,19	0,18	-0,01	20,60	5,66	-14,94
IRT3N3_5			1,40			-2,53			0,20			1,63	
IRT3N3_10			2,17			-0,15			0,22			12,64	

Table 22. N_IT3 What comes next - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N5R1_1A	IRTN1_01	0,69	0,69	0,00	-3,73	-4,67	-0,94	0,20	0,20	0,00	3,34	6,25	2,91
N5R1_2A	IRTN1_23	1,12	1,12	0,00	-0,58	-0,35	0,23	0,17	0,11	-0,06	5,90	13,92	8,02
N5R1G_3A	GRT1NR_30	0,70	0,70	0,00	0,14	-0,14	-0,28	0,27	0,14	-0,13	28,75	14,57	-14,18
N5R1_4	IRTN1_24	0,96	2,85	1,89	-1,88	-0,61	1,27	0,15	0,17	0,02	23,47	9,09	-14,38
N5R1G_5	GRT1NR_12	1,10	1,91	0,81	-1,63	-0,04	1,59	0,15	0,15	0,00	26,91	5,95	-20,96
N5R1G_6	GRT1NR_13	0,76	2,14	1,38	-1,62	-0,30	1,32	0,17	0,16	-0,01	32,49	6,71	-25,78
N5R1G_7	GRT1NR_16	0,90	1,45	0,55	-0,92	-0,47	0,45	0,17	0,18	0,01	35,54	11,05	-24,49
IRT3N3_1			0,71			-4,16			0,20			4,59	
IRT3N3_2			1,26			-2,54			0,19			3,83	
IRT3N3_4			1,33			-0,96			0,18			9,33	
IRT3N3_6			1,37			-1,74			0,18			10,90	
IRT3N3_8			2,04			-1,13			0,15			8,69	
IRT3N3_9			0,73			-1,48			0,20			6,05	

IT4 (n=296):

Table 23. N_IT4 What comes next - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N6R1_1A	IRTN1_01	0,69	0,69	0,00	-3,73	-5,05	-1,32	0,20	0,20	0,00	3,34	2,63	-0,71
N6R1_2A	IRTN1_23	1,12	1,12	0,00	-0,58	-0,34	0,24	0,17	0,10	-0,07	5,90	22,50	16,60
N6R1G_3A	GRT1NR_30	0,70	0,70	0,00	0,14	0,00	-0,14	0,27	0,13	-0,14	28,75	12,77	-15,98
N6R1G_4	GRT1NR_23	1,16	1,96	0,80	-1,35	-1,45	-0,10	0,15	0,20	0,05	27,92	13,24	-14,68
N6R1G_5	GRT1NR_26	0,90	2,20	1,30	-0,87	-0,36	0,51	0,13	0,17	0,04	23,90	10,64	-13,26
N6R1G_6	GRT1NR_27	1,43	3,54	2,11	-0,24	-0,57	-0,33	0,17	0,18	0,01	28,70	10,58	-18,12
N6R1G_7	GRT1NR_28	1,37	1,22	-0,15	-1,19	-0,22	0,97	0,14	0,22	0,08	1137,73	14,51	-1123,22
IRT3N1_1			0,97			-3,16			0,20			6,83	
IRT3N1_2			1,08			-3,21			0,20			3,01	
IRT3N1_4			2,31			-1,16			0,16			9,24	
IRT3N1_6			1,67			-1,73			0,17			6,77	
IRT3N1_8			1,29			-1,28			0,18			12,29	
IRT3N1_9			1,25			-1,16			0,20			7,70	

Table 24. N_IT4 Text-based problems - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N7R1_1A	IRTN4_16	0,75	0,75	0,00	-1,06	-0,81	0,25	0,12	0,14	0,02	18,09	21,39	3,30
N7R1_2A	IRTN5_14	0,80	0,80	0,00	-0,71	-0,26	0,45	0,13	0,17	0,04	17,85	9,19	-8,66
N7R1_3	IRTN4_07	0,58	1,50	0,92	-2,51	-1,13	1,38	0,16	0,18	0,02	33,20	5,75	-27,45
N7R1_4	IRTN4_04	0,78	2,17	1,39	-2,47	-0,87	1,60	0,18	0,16	-0,02	28,42	6,14	-22,28
N7R1_5	IRTN4_11	0,56	1,93	1,37	-2,49	-0,42	2,07	0,15	0,15	0,00	57,67	3,97	-53,70
N7R1_6	IRTN4_17	0,89	3,24	2,35	-1,76	-0,31	1,45	0,13	0,13	0,00	31,17	4,39	-26,78
N7R1_7	IRTN4_24	0,69	3,00	2,31	-1,88	-0,29	1,59	0,14	0,16	0,02	43,31	4,09	-39,22

IT5 (n=274):

Table 25. N_IT5 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N8R1_1A	IRTN33_21	1,04	1,04	0,00	-2,67	-1,98	0,69	0,20	0,18	-0,02	4,85	12,47	7,62
N8R1_2A	IRTN33_25	1,34	1,34	0,00	1,41	0,71	-0,70	0,12	0,54	0,42	9,42	10,88	1,46
N8R1_3A	IRTN33_26	1,09	1,09	0,00	0,23	1,36	1,13	0,20	0,08	-0,12	9,91	20,77	10,86
N8R1_4A	IRTN33_28	1,44	1,44	0,00	-0,53	-0,26	0,27	0,23	0,13	-0,10	9,38	14,45	5,07
N8R1_5	IRTN3_11	0,64	1,18	0,54	-2,21	-0,77	1,44	0,10	0,18	0,08	34,10	13,36	-20,74
N8R1_6	IRTN4_21	0,91	1,90	0,99	-1,20	-0,57	0,63	0,07	0,14	0,07	55,68	8,05	-47,63
N8R1_7	IRTN3_09	0,52	2,34	1,82	-0,75	-1,40	-0,65	0,11	0,20	0,09	24,37	0,75	-23,62
N9R1_5	IRTN5_06	0,89	1,63	0,74	0,85	-0,30	-1,15	0,15	0,18	0,03	41,94	5,08	-36,86
N9R1_6	IRTN5_25	1,01	2,30	1,29	0,87	0,77	-0,10	0,10	0,25	0,15	18,69	10,38	-8,31
N9R1_7	IRTN33_13	0,71	1,36	0,65	-1,82	-1,20	0,62	0,19	0,20	0,01	22,61	9,28	-13,33
N10R1_5	IRTN5_03	0,67	1,21	0,54	-1,75	0,06	1,81	0,13	0,16	0,03	46,80	7,20	-39,60
N10R1_6	IRTN33_30	1,88	1,93	0,05	0,72	0,70	-0,02	0,21	0,18	-0,03	17,32	7,46	-9,86
N10R1_7	IRTN5_20	1,11	3,82	2,71	0,10	0,40	0,30	0,11	0,09	-0,02	-	7,28	-
N1R1_5	IRTN3_04	0,91	1,32	0,41	-3,55	-0,80	2,75	0,18	0,18	0,00	23,72	10,07	-13,65
IRT3N2_11			1,96			-0,93			0,13			17,88	
IRT3N2_16			3,07			-1,00			0,17			11,57	

IT6 (n=279):

Table 26. N_IT6 Odd one out - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N11R1_1A	IRTN2_06	0,57	0,57	0,00	-3,19	-3,99	-0,80	0,18	0,20	0,02	14,87	7,24	-7,63
N11R1_2A	IRTN2_17	1,08	1,08	0,00	1,22	1,86	0,64	0,16	0,18	0,02	16,57	8,03	-8,54
N11R1_3A	IRTN2_20	1,20	1,20	0,00	0,24	0,92	0,68	0,09	0,09	0,00	12,51	15,52	3,01
N11R1G_4A	GRT2NR_17	1,86	1,86	0,00	-0,15	0,38	0,53	0,17	0,10	-0,07	18,68	7,40	-11,28
N11R1_5	IRTN5_08	0,59	1,53	0,94	1,68	-0,24	-1,92	0,16	0,16	0,00	124,79	13,90	-110,89
N11R1_6	IRTN5_10	0,58	1,26	0,68	-0,80	0,09	0,89	0,25	0,16	-0,09	34,68	8,26	-26,42
N11R1G_7	GRT2NR_10	0,78	1,39	0,61	1,46	-0,04	-1,50	0,11	0,21	0,10	9,01	12,96	3,95
IRT3_3_10			1,63			0,19			0,12			10,75	
IRT3_3_12			2,78			-0,67			0,15			4,77	
IRT3_3_14			3,22			0,29			0,13			3,14	

Table 27. N_IT6 What comes next - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
N12R1_1A	IRTN1_01	0,69	0,69	0,00	-3,73	-4,35	-0,62	0,20	0,20	0,00	3,34	8,20	4,86
N12R1_2A	IRTN1_23	1,12	1,12	0,00	-0,58	-0,27	0,31	0,17	0,11	-0,06	5,90	19,21	13,31
N12R1G_3A	GRT1NR_30	0,70	0,70	0,00	0,14	0,15	0,01	0,27	0,16	-0,11	28,75	29,92	1,17
N12R1G_4	GRT1NR_17	0,82	1,53	0,71	-1,41	-0,65	0,76	0,10	0,16	0,06	38,52	10,53	-27,99
N12R1G_5	GRT1NR_19	0,95	1,29	0,34	-1,38	0,61	1,99	0,16	0,15	-0,01	30,25	20,21	-10,04
N12R1G_6	GRT1NR_20	1,36	1,94	0,58	-1,60	-0,23	1,37	0,23	0,13	-0,10	27,02	17,33	-9,69
N12R1G_7	GRT1NR_29	1,02	1,64	0,62	-0,82	0,06	0,88	0,18	0,15	-0,03	31,90	16,79	-15,11
IRT3_3_8			1,42			-1,24			0,20			11,20	
IRT3_3_9			1,41			-0,86			0,19			19,33	
IRT3_3_13			3,53			0,18			0,15			16,66	
IRT3_3_15			2,87			0,20			0,13			23,13	
IRT3_3_17			2,99			0,80			0,06			20,62	

Similarly to the Verbal Ability parameter estimates, the Tables depicting the estimates for Numerical Ability showed improved item fit and functioning. The IRT3 items also performed effectively with difficulty values (b parameter) typically on the easier side.

3.3.4.3 Abstract Ability revised estimates

Tables 28 to 34 provide the parameter estimates for the revised Abstract items that were adapted from previously existing AdaptGRT items (IT1 to IT4) as well as IRT3 items making use of the same item format.

The new items that were developed for the Abstract Ability subtest are depicted in Table 35 (IT5) and Table 36 (IT6). These items make use of two new formats and thus it was not possible to include anchor items into the respective item sets.

The revised items (Tables 28 to 34) showed far better fit indices, indicating better functioning items overall.

IT1 (n=247):

Table 28. A_IT1 What comes next - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
A1R1_1A	IRTA1_08	1,24	1,24	0,00	-2,10	-1,82	0,28	0,13	0,18	0,05	88,48	4,43	-84,05
A1R1_2A	IRTA1_19	1,22	1,22	0,00	0,32	1,29	0,97	0,06	0,06	0,00	37,14	6,35	-30,79
A1R1_3A	IRTA1_05	0,93	0,93	0,00	1,26	1,70	0,44	0,20	0,15	-0,05	58,19	8,21	-49,98
A1R1_4	IRTA1_01	0,59	2,56	1,97	-0,69	-1,48	-0,79	0,34	0,17	-0,17	113,33	5,02	-108,31
A1R1_5	IRTA1_03	0,51	1,57	1,06	-2,18	-1,11	1,07	0,21	0,16	-0,05	197,10	10,79	-186,31
A1R1_6	IRTA1_14	1,30	2,04	0,74	1,41	-0,55	-1,96	0,03	0,12	0,09	495,54	10,97	-484,57
A1R1_7	IRTA1_17	1,07	2,53	1,46	0,08	-1,04	-1,12	0,26	0,17	-0,09	2194,97	3,38	-2191,59
IRT3A1_2			2,65			-0,52			0,19			10,33	
IRT3A1_6			1,57			0,71			0,18			9,77	
IRT3A1_8			2,24			-0,71			0,16			17,57	
IRT3A1_11			1,64			-0,54			0,18			7,53	
IRT3A1_13			1,97			0,48			0,13			4,36	

IT2 (n=312):

Table 29. A_IT2 What comes next - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
A2R1_1A	IRTA1_08	1,24	1,24	0,00	-2,10	-2,51	-0,41	0,13	0,19	0,06	88,48	4,51	-83,97
A2R1_2A	IRTA1_19	1,22	1,22	0,00	0,32	0,84	0,52	0,06	0,08	0,02	37,14	10,14	-27,00
A2R1_3A	IRTA1_05	0,93	0,93	0,00	1,26	1,66	0,40	0,20	0,09	-0,11	58,19	10,56	-47,63
A2R1_4	IRTA5_01	0,60	1,14	0,54	-1,97	-0,50	1,47	0,17	0,18	0,01	143,85	5,00	-138,85
A2R1_5	IRTA5_03	0,58	0,78	0,20	-2,00	-0,84	1,16	0,20	0,20	0,00	128,89	10,50	-118,39
A2R1_6	IRTA5_08	0,75	1,58	0,83	-2,64	-1,97	0,67	0,17	0,20	0,03	130,69	4,82	-125,87
A2R1_7	IRTA1_09	0,90	3,16	2,26	-0,84	-0,62	0,22	0,10	0,13	0,03	85,55	1,46	-84,09
IRT3A2_2			2,20			-1,11			0,18			3,95	
IRT3A2_6			3,88			0,71			0,24			7,42	
IRT3A2_8			2,92			-1,07			0,17			4,72	

Table 30. A_IT2 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
A3R1_1A	IRTA4_13	0,80	0,80	0,00	-1,86	-5,88	-4,02	0,21	0,20	-0,01	78,93	^	^
A3R1_2A	IRTA4_19	1,32	1,32	0,00	0,01	0,36	0,35	0,10	0,11	0,01	22,57	13,15	-9,42
A3R1_3A	IRTA2_10	1,17	1,17	0,00	1,51	-3,32	-4,83	0,13	0,20	0,07	33,64	5,54	-28,10
A3R1_4	IRTA2_02	0,61	3,19	2,58	-1,84	-2,04	-0,20	0,17	0,18	0,01	107,97	3,71	-104,26
A3R1_5	IRTA2_05	0,59	2,02	1,43	-3,05	-2,96	0,09	0,17	0,20	0,03	148,43	^	^
A3R1_6	IRTA2_06	0,48	1,40	0,92	0,87	-1,81	-2,68	0,25	0,19	-0,06	131,14	3,86	-127,28
A3R1_7	IRTA2_17	0,42	1,88	1,46	0,47	-2,27	-2,74	0,22	0,19	-0,03	143,27	10,50	-132,77
IRT3A2_3			1,31			0,24			0,16			1,01	
IRT3A2_7			2,13			-0,98			0,20			2,51	
IRT3A2_10			1,63			-2,67			0,20			6,56	

IT3 (n=316):

Table 31. A_IT3 Odd one out - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
A4R1_1A	IRTA3_09	1,14	1,14	0,00	0,36	-0,06	-0,42	0,23	0,24	0,01	15,16	7,14	-8,02
A4R1_2A	IRTA5_04	0,74	0,74	0,00	0,22	0,79	0,57	0,17	0,08	-0,09	89,72	17,81	-71,91
A4R1_3A	IRTA5_10	0,92	0,92	0,00	0,20	0,37	0,17	0,10	0,20	0,10	41,07	6,86	-34,21
A4R1_4	IRTA4_06	0,59	1,15	0,56	-1,39	-2,24	-0,85	0,25	0,19	-0,06	103,55	8,02	-95,53
A4R1_5	IRTA4_21	0,65	1,36	0,71	-1,43	-0,36	1,07	0,18	0,19	0,01	93,53	3,48	-90,05
A4R1_6	IRTA5_25	0,53	0,34	-0,19	-0,33	-4,99	-4,66	0,13	0,20	0,07	106,78	19,76	-87,02
A4R1_7	IRTA3_21	0,67	0,98	0,31	0,35	-3,50	-3,85	0,18	0,20	0,02	33,55	2,45	-31,10
IRT3A3_1			0,36			1,03			0,20			7,93	
IRT3A3_4			1,75			-0,82			0,17			8,68	
IRT3A3_5			1,16			0,44			0,23			12,36	
IRT3A3_9			0,94			-0,01			0,21			7,91	

Table 32. A_IT3 XYZ - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
A5R1_1A	IRTA4_13	0,80	0,80	0,00	-1,86	-5,06	-3,20	0,21	0,20	-0,01	78,93	5,62	-73,31
A5R1_2A	IRTA4_19	1,32	1,32	0,00	0,01	-0,08	-0,09	0,10	0,11	0,01	22,57	5,40	-17,17
A5R1_3A	IRTA2_10	1,17	1,17	0,00	1,51	-3,28	-4,79	0,13	0,19	0,06	33,64	9,62	-24,02
A5R1_4	IRTA4_20	0,55	0,89	0,34	-0,91	-1,20	-0,29	0,17	0,19	0,02	105,62	5,86	-99,76
A5R1_5	IRTA5_02	0,74	1,79	1,05	-3,24	-2,35	0,89	0,19	0,20	0,01	124,96	3,55	-121,41
A5R1_6	IRTA2_20	0,80	1,61	0,81	-1,33	-1,82	-0,49	0,20	0,19	-0,01	60,99	4,83	-56,16
A5R1_7	IRTA2_12	0,52	0,98	0,46	1,48	-0,37	-1,85	0,14	0,18	0,04	65,39	2,81	-62,58
IRT3A3_3			1,24			-1,05			0,19			4,76	
IRT3A3_7			1,64			-1,03			0,16			1,98	
IRT3A3_10			2,98			-1,43			0,20			5,86	

IT4 (n=307):

Table 33. A_IT4 What comes next - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
A6R1_1A	IRTA1_08	1,24	1,24	0,00	-2,10	-2,27	-0,17	0,13	0,18	0,05	88,48	12,13	-76,35
A6R1_2A	IRTA1_19	1,22	1,22	0,00	0,32	0,64	0,32	0,06	0,06	0,00	37,14	18,04	-19,10
A6R1_3A	IRTA1_05	0,93	0,93	0,00	1,26	1,34	0,08	0,20	0,11	-0,09	58,19	8,74	-49,45
A6R1_4	IRTA1_29	0,88	0,78	-0,10	-1,53	-5,08	-3,55	0,10	0,20	0,10	79,37	1,43	-77,94
A6R1_5	IRTA1_30	0,80	1,84	1,04	-2,63	-2,78	-0,15	0,18	0,20	0,02	702,51	1,80	-700,71
A6R1_6	IRTA1_12	1,07	1,02	-0,05	-0,89	-1,05	-0,16	0,13	0,19	0,06	63,07	5,45	-57,62
A6R1_7	IRTA1_16	0,54	3,21	2,67	0,71	0,17	-0,54	0,07	0,12	0,05	43,04	4,39	-38,65
IRT3A1_2			2,07			-1,04			0,21			7,65	
IRT3A1_6			1,95			0,74			0,14			5,31	
IRT3A1_8			1,70			-1,59			0,18			9,44	

Table 34. A_IT4 XYZ   - 3PL Model Item Parameter Estimates.

Trial Name	Original name	Original a	Revised a	Change a	Original b	Revised b	Change b	Original c	Revised c	Change c	Original fit	Revised fit	Change fit
A7R1_1A	IRTA4_13	0,80	0,80	0,00	-1,86	-5,48	-3,62	0,21	0,20	-0,01	78,93	5,58	-73,35
A7R1_2A	IRTA4_19	1,32	1,31	-0,01	0,01	0,46	0,45	0,10	0,10	0,00	22,57	4,72	-17,85
A7R1_3A	IRTA2_10	1,17	1,17	0,00	1,51	-3,56	-5,07	0,13	0,20	0,07	33,64	5,04	-28,60
A7R1_4	IRTA2_24	0,76	0,95	0,19	-1,71	-3,59	-1,88	0,13	0,20	0,07	121,02	7,81	-113,21
A7R1_5	IRTA4_11	0,97	1,15	0,18	-0,30	-1,86	-1,56	0,09	0,20	0,11	391,54	4,98	-386,56
A7R1_6	IRTA4_12	1,01	1,35	0,34	1,18	-0,66	-1,84	0,15	0,17	0,02	58,02	3,93	-54,09
A3R1_4	IRTA2_02	0,61	1,53	0,92	-1,84	-2,62	-0,78	0,17	0,20	0,03	107,97	4,53	-103,44
IRT3A1_3			1,32			0,01			0,16			7,56	
IRT3A1_7			1,78			-0,39			0,22			6,78	
IRT3A1_10			1,97			-1,63			0,19			11,70	

IT5 (n=294):

Table 35. A_IT5 Connector - 3PL Model Item Parameter Estimates.

Trial item name	a	b	c	fit
A8R1C_1	2,37	-1,03	0,16	11,10
A8R1C_31	1,91	-0,4	0,16	5,28
A8R1C_20	2,97	-0,75	0,23	6,56
A8R1C_14	3,08	-0,81	0,21	6,93
A8R1C_62	3,13	-0,21	0,13	7,42
A8R1C_69	0,68	0,38	0,20	8,87
A8R1C_51	1,62	-0,09	0,14	9,21
A9R1C_4	3,78	-1,25	0,17	8,24
A9R1C_23	1,96	-1,11	0,16	4,93
A9R1C_54	1,89	-0,36	0,23	7,25
A9R1C_16	3,18	-1,08	0,15	6,28
A9R1C_35	1,37	0,07	0,16	7,37
A9R1C_66	2,4	0,46	0,18	5,86
A9R1C_71	1,29	-0,16	0,19	10,43

IT6 (n=320):

Table 36. A_IT6 Puzzle - 3PL Model Item Parameter Estimates.

Trial item name	a	b	c	fit
A10R1P_1	1,51	-2,47	0,20	13,23
A10R1P_37	1,01	-1,57	0,19	10,32
A10R1P_43	2,34	-2,02	0,20	3,76
A10R1P_32	1,1	-0,23	0,20	8,34
A10R1P_14	1,84	-0,88	0,13	18,82
A10R1P_27	0,62	0,38	0,21	5,69
A10R1P_21	1,57	-0,09	0,26	10,93
A11R1P_46	2,92	-1,37	0,19	5,25
A11R1P_17	1,2	-0,62	0,16	6,97
A11R1P_24	2,12	0,02	0,18	4,99
A11R1P_10	1,66	-1,37	0,20	7,56
A11R1P_29	1,11	-0,66	0,19	6,04
A11R1P_36	1,41	-0,59	0,17	14,53
A11R1P_41	0,97	0,19	0,19	3,29

3.4 DESIGN PHILOSOPHY AND ON-GOING ANALYSIS

Wave three is conceptualised as a comprehensive and on-going approach that incorporates item calibration, deletion or revision, seeding, trialling and recalibration. The goal is to remove problematic items, revise those that are potentially confusing and broaden the item pool to address gaps in the trait continuum. In this sense, the design philosophy is that of a test that is never finished, but that is continually refined over time with improved and new items, and updated parameter estimates.

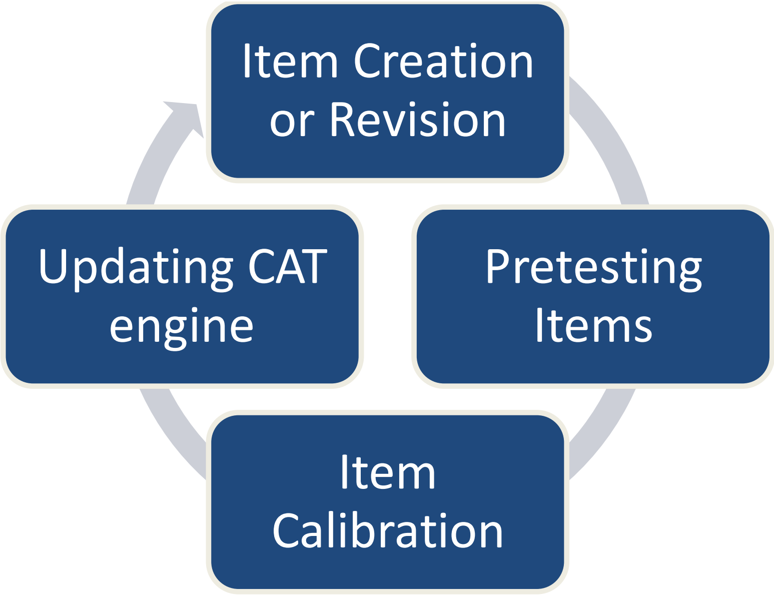
The parameter and fit estimates of existing items are checked periodically and potentially problematic items are revised. New items within existing formats and new formats are also continually being developed in parallel to the revision process to ensure that the item pool contains items that discriminate across the trait continuum at close intervals and with maximum accuracy

New or revised items are pretested by seeding them into the AdaptGRT. The test is taken under high stake conditions but the trialled items do not contribute toward an examinee's score or test performance.

Once a sufficient number of responses has been gathered, each subset of items is calibrated using the IRTPRO programme. If found to have acceptable parameter estimates and model fit, these items are formally incorporated into the assessment. If not, the items are reviewed for deletion or revision. The cultural sensitivity of the items is assessed to ensure that they discriminate equally between individuals from different cultures. This process will continue in a cyclical manner until analysis shows satisfactory statistics with low standard error, medium to high IIFs and no evidence of multicollinearity.

The design philosophy of wave three is depicted in figure 12.

Figure 12. Wave Three Design Philosophy.



REFERENCES

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*(7), 565-580.

Baker, F. B. (2008). The basics of item response theory. ERIC Clearinghouse on Assessment and Evaluation, 2001.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397 – 472). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*(4), 431-444.

Brown, G.S., & McInnes, A. (2011). *AdaptGRT: Test-Retest Reliability*. Unpublished research report.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software].*Chicago, IL: Scientific Software International*.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York, NY: Cambridge University Press.

Cattell, R. B. (1971). Abilities: Their structure, growth, and action. Boston, MA: Houghton Mifflin.

Du Toit, M. (Ed.). (2003). *IRT from SSI: Bilog-MG, multilog, parscale, testfact*. Scientific Software International.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous item response theory models to multiple-choice tests.*Applied Psychological Measurement*, *19*(2), 143-166.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological review*, *75*(3), 242.

Hunter, J. E. (1980). Validity generalization for 12,000 jobs: An application of synthetic validity and validity generalization to the General Aptitude Test Battery (GATB). Washington, DC: U.S. Department of Labor, Employment Service

Kline, P. (2000). *A psychometrics primer*. Free Assn Books.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. Springer New York.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive metaanalysis of the predictive validity of the Graduate Record Examination: Implications for graduate student selection and performance. Psychological Bulletin, 127, 162-181.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? Journal of Personality and Social Psychology, 86, 148-161.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement*, *23*(3), 187-194.

Psytech International (2010). *General and graduate reasoning tests.* Unpublished technical manual. Available: http://www.psytech.com/Manuals/GRT2Man.pdf

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Schneider, W. J. , & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. Flanagan & P. Harrison (Eds.), Contemporary intellectual assessment: Theories, tests, and issues (3rd ed., pp. 99–144). New York: Guilford.

Spearman, C. (1904). “General intelligence,” objectively determined and measured. American Journal of Psychology, 15, 201-293.

Thissen, D., & Mislevy, R.J. (2000). Testing Algorithms. In Wainer, H. (Ed.) Computerized Adaptive Testing: A Primer. Mahwah, NJ: Lawrence Erlbaum Associates.

Unick, G. J., Shumway, M., & Hargreaves, W. (2008). Are we ready for computerized adaptive testing? *Psychiatric services (Washington, DC)*, *59*(4), 369.

Van Der Linden, W., & Hambleton, R. (1997). Handbook for Modern Item Response Theory. New York: Springer.

Van der Linden, W. J., & Glas, C. A. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education*, *13*(1), 35-53.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 361-375.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, *21*(1).