Implementing Computer Adaptive Testing to Improve Achievement Opportunities

by Michael Birdsall.

Ofqual/11/4859 April 2011

Ofqual Office of Qualifications and Examinations Regulation

This report has been commissioned by the Office of Qualifications and Examinations Regulation.

1. Executive Summary

This paper is a consultation prepared for Ofqual that outlines the process of developing a computer adaptive test with special attention given to methods used to mitigate test and question biases that impact members of different population subgroups.

The scope of this paper does not include key stages 1, 2 and 3. Although there are CAT programs administered to this age group in the United States, the validity of these exams has not been thoroughly researched. In some cases, exams administered to this population may have no predictive validity due to erratic learning curves of young children. Any CAT program intended for this population would likely need further research.

This paper has three sections:

- A general overview of how computer adaptive testing works.
- A review of the development of computer adaptive tests as experienced by psychometricians in the United States, in the attempt to develop a set of best practices that coincide with practices recommended to the UK awarding organizations in Wheadon, Whitehouse, Spalding, Tremin, and Charman (2009), Boyle (2010), and He (2010).
- Practices are suggested to take advantage of computer adaptive testing to eliminate biased questions and biased exams that impact population subgroups.

This format was chosen because it is generally not possible to understand what benefits can be offered by a computer adaptive test, if one is not familiar with the assumptions underlying computer adaptive testing, the process of constructing a computer adaptive test, and the practices required to run a computer adaptive testing program.

Throughout this paper costs of development are given for the major expenses likely to be incurred in test development.

Recommendations are suggested for organizations wishing to pursue a computer adaptive testing program.

2. Background

2. 1 Computer Adaptive Testing.

A computer adaptive test (CAT) is an exam administered on a computer that adapts the difficulty level of each question or item to the ability level of the candidate.

There is a long history of a desire to use adaptive testing to improve testing results. The Binet-Simon intelligence test, first administered in 1905, was adaptive (van der

Linden, Glas (eds.) 2010). Under the Binet-Simon the test administrator chose questions depending on the response of the candidate. Because having an administrator available for each candidate was costly, these adaptive tests were never used on a large scale. However, such a testing program is still useful to consider as an ideal test, where each candidate receives an exam administered specifically for that candidate. Computer adaptive testing aims to move closer to that ideal.

In computer adaptive testing, the difficulty level of the test items is determined by the ratio of the number of past candidates who answered the item incorrectly to the total number of candidates who viewed the item. An item that many candidates get incorrect is determined to be difficult. An item that many candidates get correct is determined to be easy. A candidate who answers correctly items that many candidates answer incorrectly will get a higher score than a candidate who answers correctly those items that nearly all candidates answer correctly. While this may seem reasonable, it is a departure from the practice of using subject matter experts to determine the difficulty level of an item. Using computer adaptive testing models, there is no subjective measure of an items difficulty. Difficulty is strictly a statistical parameter.

The process of computer adaptive testing can be explained as a series of steps:

1. After receiving instruction in the use of the system, a candidate views the first item, which is chosen from an item bank to meet a predefined criterion.
2. If that item is answered correctly, a more difficult item will be selected from the item bank. If that item is answered incorrectly, a less difficult item will be selected from the item bank.
3. As each item is answered the candidate's provisional ability level is updated and that ability level is used to select subsequent items with a difficulty level corresponding to the candidate's provisional ability estimate.
4. The process continues until the test has met a predefined end criterion.
5. The candidate's final score is calculated.

These steps are useful for understanding the process. However, in practice, computer adaptive testing is more sophisticated compensating for such factors as balanced test content, the likelihood of candidates cheating, and items that affect subgroups of the population.

There are three primary algorithms used in computer adaptive testing:

- an item calibration algorithm,
- an item selection algorithm, and
- a candidate scoring algorithm.

The use of statistical data to determine the parameters of the items and the use of those parameters to make inferences about candidates' underlying abilities is the basis of item response theory (IRT); the psychometric paradigm behind many of the

computer adaptive tests in use today. The US based CATs reviewed for this paper both use a 3 parameter IRT model, in which there are three data points that define an item. Those data points can be thought of as

- the difficulty of a question,
- the capability of a question to differentiate between different ability levels, and
- the likelihood that a candidate would get a question correct simply by guessing.

An important distinction to clarify is that the word *ability*, as used in this paper, is more generally referred to in the existing literature as a latent trait of the candidate (de Ayala, 2009). Although He (2010) briefly mentioned latent ability, the distinction needs more clarification because no assessment, including an IRT based CAT, is capable of directly measuring all those skills that are typically being assessed.

For example, an item designed to test students' understanding of a written passage by asking a question about that passage does not actually test students' understanding. The question tests students' ability to answer questions about the passage, but not understanding. The underlying assumption is that there is a significant dependent relationship between the ability to answer questions about a reading passage and the understanding of the passage. The ability to answer the question may depend on multiple skills including the ability to read the passage, the ability to cope with the pressure of taking an exam, and the ability to correctly interpret the question. All of these separate abilities can be referred to as the latent trait the item is designed to test. While this may seem like an exercise in semantics, an understanding that a single item may test multiple abilities is particularly important when attempting to eliminate bias against subgroups in examinations. As mentioned above, throughout this paper, *ability* should be understood as a latent trait, i.e. those skills necessary to correctly answer an item, not as the distinct ability or a single skill.

Latent trait directly relates to questions of item and test validity. The ability being tested on any exam is the ability to do well on that exam. If the exam is valid, there is a strong relationship between a candidate's exam results and the candidate's ability in the area the exam was designed to test. For example, a Math GCSE tests a candidate's ability to take a Math GCSE. If the exam is valid, a regression analysis of a candidate's performance on a Math GCSE and a candidate's math skills would suggest a causal relationship between those math skills and the Math GCSE results. If an exam is a valid measure of the latent trait, then a similar regression analysis between test items and an exam would be used to demonstrate the validity of an item. Just as the difficulty parameter is a statistical measure so too is validity.

With this understanding, it is important to acknowledge that all existing IRT based CATs are unidimensional. Every item on an exam measures the ability of a candidate to perform on that exam. In the example used above, we could say that every item on a Math GCSE is an assessment of a candidate's ability on a Math GCSE. It is important not to confuse the unidimensionality of an exam with the content of an exam. The

content of the GCSE is determined by the specification. Each of the content areas addressed in the specification are indicators of a candidate's ability in Math GCSE. Since the specification of a Math GCSE requires several content areas be tested, any computer adaptive test will need to address methods of ensuring all of the content areas are tested. This is referred to as balancing content. This paradigm could be used to explain current exam practices in the UK and is not unique to IRT based computer adaptive tests. Because candidates are not given scores in each of the content areas of a GCSE, it is fair to say that unidimensionality exists on current GCSEs to the same extent that it would exist if the pencil and paper tests were replaced with CATs. While there is promising research in multidimensional item response theory and there have been successful tests of multidimensional computer adaptive tests (van der Linden, *et al.,* 2010), there are currently no operating multidimensional computer adaptive tests being used in high stakes testing.

A clear understanding of latent trait theory, exam and item validity, and unidimensionality will aid in complying with the Regulatory Principles for e-Assessment section 1.1 (QCA, 2007).

For more technical information on IRT please see de Ayala (2009), Baker and Kim (2004), He (2010), and van der Linden et al. (2010).

2.2 Computer Adaptive Testing in the United States

The ASVAB (Armed Services Vocational Aptitude Battery) was the first high stakes computer adaptive test to be implemented (Segall and Moreno, 1999). It was a paper and pencil test that underwent the transition to a computer adaptive test. It tests both vocational and academic skills, is administered to more than 1 million candidates annually (Pommerich, Segall, and Moreno, 2009), and, since 2008, has been administered over the Internet. Due to its scale and the amount of research supporting it, the ASVAB serves as a good reference for those implementing a computer adaptive test.

The GMAT® (Graduate Management Admission Test) is administered world wide at more than 400 testing centers to more than 200,000 candidates annually (Rudner, 2007). The GMAT® is used as part of the admissions process to graduate management programs. Due to its global administration, the test serves as a good reference for a flexible and culturally sensitive testing program.

Other exams will be mentioned in this paper but only for one characteristic of the exam that is used to eliminate item or test bias. The ASVAB and GMAT® will be referred to for the development of a CAT.

While other United States based tests, such as the SAT® Reasoning Test, have been reviewed by researchers in the UK (Wheadon *et al.*), the SAT® is not computer adaptive and has not been considered.

There are very successful computer adaptive programs outside the United States. For example the Psychometric Entrance Test administered in Israel (Gafni, N., Cohen, Y., Roded, K., Baumer, M., & Moshinsky, A., 2009), the Multiple Choice Exam (MCQ) administered in Australia to candidates hoping to attend medical school, and the Medical Council of Canada's Qualifying Examination Part 1 (MCCQE Part 1) administered in Canada. The inclusion of only United States based computer adaptive testing programs does not suggest that only United States based practices need be considered when defining best practices, but rather that a good place to start is by examining some of the most well researched computer adaptive programs. The GMAT® and the ASVAB were selected for this reason.

3. Test Development.

Thompson and Weiss (2011) outline a framework for the development of a computer adaptive test. That framework consists of five steps and mentions a sixth included below:

1. A feasibility study.
2. Item bank development.
3. Pre-testing and analyzing the items.
4. Taking decisions on test specifications.
5. Publishing the computer adaptive test.
6. Maintaining the test.

Since a computer adaptive test requires continuous development it is best think of these steps as the cycle illustrated by Figure 1 below.

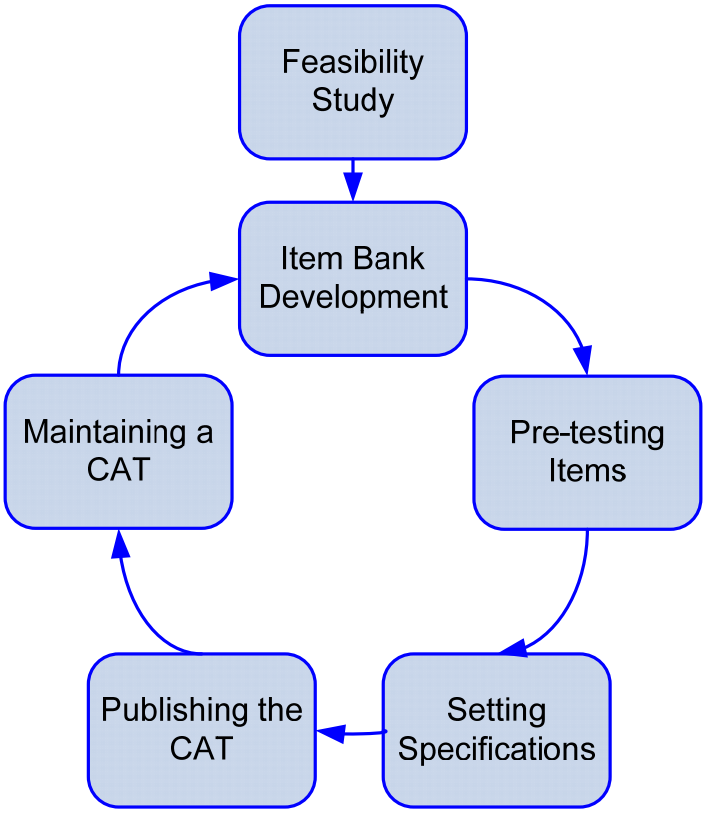


Figure 1.

In general, this framework is consistent with the implementation of ASVAB as documented by Wise, Curran, and McBride (1997), Segall *et al.* (1999), and Pommerich *et al.* (2009). The framework is also consistent with the implementation of the GMAT® as documented by Rudner (2007).

3.1 Feasibility Study

3.1.1 Benefits of Computer Adaptive Testing

Computer adaptive testing has been replacing paper and pencil tests for a number of reasons.

1. Results can be more reliable than paper and pencil results.[1] (Pommerich *et al.*, 2009)
2. Greater reliability results in more satisfying placements of candidates (Segall *et al.,* 1999).
3. Computer adaptive tests generally require less time to administer and therefore save money (Pommerich *et al.*, 2009).
4. CATs can be administered on demand at several locations increasing accessibility (Rudner, 2007).
5. CATs increase the statistical accuracy of an assessment (van der Linden et al., 2010)

Wheadon *et al.* (2009) pointed out that on-demand testing can also increase efficiency by

- providing immediate feedback to students, teachers and others,
- improving flexibility by facilitating testing when teachers and students are ready for exams, and
- increasing data flow between stakeholders.

Further, Boyle (2010) cites increased motivation as a result of computer adaptive testing.

All of the factors above can result in cost savings for multiple stakeholders. Thus, if the decision is made to change to such a testing system, the question of who should bare the cost of converting does not have a clear answer.

3.1.2 Obstacles to Computer Adaptive Testing

In addition to those obstacles outlined by the Thomson Report entitled Drivers and Barriers to the adoption of e-Assessment for UK Awarding organizations, four obstacles face the awarding organizations:

- Most computer adaptive tests use a multiple choice format, whereas the GCSEs and A-Levels have short answer and essay items.

[1] Pommerich *et al.* (2009) reports that the correlation between score on the CAT-ASVAB and two separate pencil and paper exams was higher than the correlation between those two pencil and paper exams.

- The awarding organizations may not have sufficient experience and psychometric expertise in house to develop CATs (Wheadon *et al.,* 2009 and He, 2010).
- The lack of expertise and understanding of CAT may result in objections to the use of CAT testing that stem from a superficial understanding of IRT and/or CAT.
- The GCSE and A-Levels test multiple subjects. Development of computer adaptive tests for all subjects may be very expensive.

The decision of whether to stick with short answer and essay or move to multiple choice items is a decision that should be made early on.

It is possible to construct a computer adaptive test that uses a linguistic analysis of a candidate's responses to choose the next question. In this model, the provisional test score, as calculated by linguistic features, would be used for selecting items, but the final score would be calculated afterwards by a human reader. The essay section of the GMAT® is marked by software, and has been reported to have an average agreement with human readers of 97% (Rudner, Garcia, & Welch, 2006). Thus it may not be necessary to use human readers at all. However, from a public relations point of view, it is best to continue using human readers.

The essay section of the GMAT® is not computer adaptive. To make it computer adaptive, a two parameter partial credit IRT model could be adopted. The partial credit received for each item would be determined by linguistic features of the answer.

It may seem as though awarding organizations would need to choose between multiple choice and essay questions. However, if the same population of students were to complete both essay and multiple sections those results could be linked. The benefits of linking are that those items that performed very differently between the two types of tests could be analyzed for biases that originate from the different question types.

Although this is theoretically possible, this use of linguistic readers needs further research. Such research will be required if the current format of short answer and essay items is to be retained. The cost of research could likely be lowered by working with vendors to get free training and software in agreement for conducting the research. There are many benefits to vendors of having a computer adaptive essay grader proven in the market. Thus vendors have a strong motivation to carry out the research at minimal cost.

There are however benefits of moving to a multiple choice model. As Wheadon *et al.* (2009), pointed out, "multiple choice obviates appeals." Since items on CATs have been pretested, there is a very small likelihood that a scored question on an exam will not be valid. The resulting cost savings obtained by obviating appeals would have to be balanced by the need to maintain trust in the system.

Publications from 2009 and 2010 (Wheadon *et al.*, 1999*,* Boyles, 2010, and He, 2010) suggest that there may not be enough experience with IRT-based CATs within the awarding organizations to implement a successful CAT program at present. The same literature suggested that some collaboration between awarding organizations may be useful. In spite of these recommendations, potential competition between bodies suggests it may be wise for each awarding organization to develop their own program. This lack of expertise raises two questions:

1. If awarding organizations do not have the expertise to develop CATs, who does?
2. If awarding organizations do not have the expertise to develop CATs, do they have the expertise to evaluate the benefits of a CAT program?

The answer to the first question may be that higher education bodies in the UK already have the expertise required. For example, the UK clinical aptitude test or UK CAT is a CAT program used for assessing applicant's likelihood of success as a medical student (Wu, 2010). So while awarding organizations may not have the expertise, may be readily available.

If no collaboration with higher education institutions is possible, awarding organizations may develop a CAT program using the same methods used by CAT developers in the United States.

The experiences of the GMAT®'s developers and the ASVAB's were different. Primarily because the ASVAB was the first successful high stakes CAT, there was no opportunity to bring in outside consultants. Thus the ASVAB's developers underwent three stages of study at the feasibility stage of product development: theoretical analyses, simulation studies, and empirical studies. The empirical studies consisted of a multiyear pilot program at six testing sites, which led to increased confidence in the system among stakeholders (Pommerich *et al.*, 2009). While the GMAT®'s developers were able to employ outside consultants, it is notable that doing so only eliminated the theoretical analysis stage of feasibility. However, the GMAT® did make the conversion from paper and pencil to CAT over a shorter period of time than did the ASVAB.

Since an empirical study is necessary, and performing that study as pilot programs may also raise confidence in the program among stakeholders, an awarding organization hoping to implement a CAT would benefit from running a pilot program.

If an awarding organization does not have the skills necessary to implement a CAT, it is likely that it will need to retain consultants. When the GMAT®'s developers took this approach, they also took on an independent psychometrician to evaluate the advice of the consultants. If a similar program is implemented, skills transfer from both the consultant team and the third party psychometrician may be desirable. For example, the awarding organization could choose to implement the GCSE for Biology as a pilot using outside consultants. Employees of the awarding organizations would work with

the consultancy team, under the supervision of the in-house psychometrician. Upon successful implementation of the Biology GCSE, those employees of the awarding organizations would implement the Chemistry and Physics GCSE, while training more employees. Such a gradual approach would have the benefit of developing in-house expertise, which would be used for scaling up the program at an exponential speed, building stakeholder confidence, and, of course, ensuring that a quality assessment is produced.

Many of the steps needed to develop a CAT will need to be repeated to maintain the system. Even if the training program is not implemented as described above, awarding organizations will eventually need to ensure they have those skills in-house. As the awarding organization increases the number of CAT assessments offered, the opportunity to realize economies of scale, would suggest that outside consultants are perhaps a good temporary strategy.

Due to the current pressures on schools in the United States, psychometricians are expensive to hire. Total benefit packages for psychometricians range from 100,000 USD annually to 350,000 USD annually. Indeed psychometricians in the psychometrics center at Cambridge University quote a rate of 100,000 GBP a year. Once training, or hiring an in house psychometirican, awarding organizations will need to offer similar pay packages adding to the cost of maintaining an ongoing CAT program.

If awarding organizations do not have the expertise available to evaluate whether they should implement a computer adaptive program, the barrier to the adoption of a computer adaptive program is significant. Without direct government intervention, such a barrier could last a generation.

Planning for the training of in house staff will meet section 11.1 of Regulatory Principles for e-Assessment.

3.1.3 Modeling Costs

After taking a decision on the format of test items and the pilot subject, statistical software can be used to run a simulation to determine the cost of switching to a computer adaptive test (Thompson *et al.*, 2011).

The statistical software would determine how many additional items awarding organizations would need to bank to be able to achieve a desired reliability level on a CAT.

When estimating costs, awarding organizations should not overestimate cost savings due to shorter test administration times. The most frequent accommodation granted is

extra time. As a result, not considering accommodation in cost estimates could result in inaccurate modeling of costs.

Wheadon *et al.* (2009) pointed out that the number of retakes may increase as a result of on-demand testing. Such an increase could increase revenues for the awarding organizations. As there will be many stakeholders benefiting from the transition to CAT, this increased revenue should be modeled as it may help to share the cost of development among stakeholders.

3.2 Item Bank Development

Based on the results of the simulation, new items will need to be written. If an exam is to be reliable at all levels of the ability scale, the item bank should contain an adequate number of items at every level of the difficulty scale.

Item bank development starts with the specification. The domains of a specification contain multiple tasks that are thought to demonstrate the ability being tested. For example, on the AQA GCSE Math specification the first five major domains mentioned are

- working with numbers and the number system,
- fractions, decimals, and percentages,
- ratio and proportion,
- the language of algebra, and
- sequences, functions, and graphs (AQA, 2009).

Under the domain of ratio and proportion, three tasks are explained that are thought to demonstrate a candidate's ability in this domain. These three tasks are

- use ratio notation, including reduction to its simplest form and its various links to fraction notation,
- solve problems involving ratio and proportion, including the unitary method of solution, and
- repeated proportional change. (The standard notes that this is a higher order skill.)

See Figure 2 below.

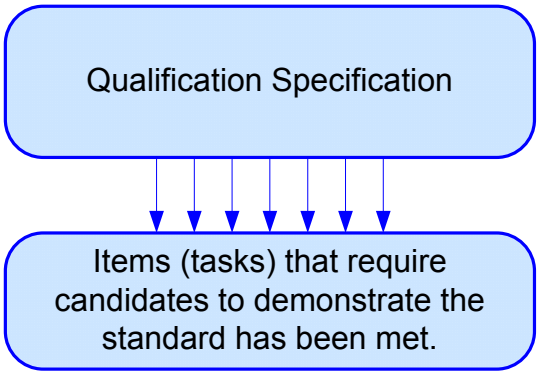


Figure 2.

It is neither feasible nor necessary to assign each task in each content area to every candidate. In practice a candidate is given one or two of the tasks outlined above. One of the concerns of CAT developers is to ensure that each candidate receives a proportional number of tasks from each content area. As an extreme example, consider the possibility that one candidate received an exam with only fraction decimal and percentage questions, while another candidate received only geometry questions. Would the specification have been met? Would the exam results be comparable?

This concern is addressed by the practice of content balancing on the exam. To balance an exam, new items representing a variety of tasks may have to be written for each of the content areas tested in the exam. Pommerich *et al*. (2009) reports that there are a proportional number of items used in each content domain on the ASVAB. Another approach used is to tag items with the subject and the domain, or domains, tested by the item. The item selection algorithm will then ensure each candidate is given items that contain the same content even though the exams will be different. Adding any constraints to an item selection algorithm, will result in the need for the size of the item bank to be increased. New banks of questions will be assigned to separate pools available for different test administrations. Construction of those pools should not be completed while writing the items, but the characteristics of each pool will determine how many items must be in the bank.

Rudner (2007) noted that the cost of developing the GMAT® CAT was 11.7 million USD and that a large portion of that bill was attributable to item development, with each item costing between 1500 and 2500 USD to develop. It was noted that CAT items can have a very long life.

3.3 Pretesting the items.

After items have been developed, they should be tested in exam-like conditions. Doing so will meet the requirement of Regulatory Principles for e-Assessment section 5.1.

3.3.1 Seeding Items

Wheadon *et al.* (2009) states:

> "Pretesting in live tests provides the highest level of quality assurances for those who set the tests and evaluate those tests."

A common method of pretesting items is to seed them into an existing exam. Seeding is the practice of placing items that do not count toward the candidates score in a live test. Pommerich *et al.* (2009) notes that there is evidence that comparable results can be achieved by seeding even if the mode of testing is different. However, in practice, both the ASVAB's and GMAT®'s developers pretested items used in operational CAT's on computer adaptive tests. While the ASVAB's developers had the benefit of using military service members to pretest, the GMAT®'s developers offered pilot test candidates the option of keeping the better of the two scores.

Seeding is a popular strategy because it ensures that candidates have the same motivation while answering pre-tested items as they would while answering real items. He (2010) points out a criticism of seeding; it could result in two different testing experiences. In practice, those items may not affect a candidate's result. The ASVAB seeds one item per exam (Pommerich *et al.*, 2009). The effect could be further mitigated by rescoring the exam once the seeded item is calibrated. In this scenario, only those items that have no further need for development would be used in scoring. So if one third of the seeded items needed further development, two thirds of candidates could have their exams rescored. Since one unscored item is unlikely to greatly affect a candidate's score, this is largely an area for further research as there is not much interest in studying a practice that is unlikely to impact candidates' scores.

3.3.2 Analysis of the Data.

After the items have been administered, the collected data will be analyzed to ensure the scale of the exam is comparable with previous administrations of the same exam, ensuring the data fit the item response theory model, ensuring that only the subject anticipated is being tested by the item, and ensuring the data collected is the same for all subgroups of the population.

In order to provide stakeholders with a familiar scoring system (e.g. grades from A to E) most programs retain the existing presentation of results. However, such systems are by their nature arbitrary.

The ASVAB uses a straight forward percentile ranking. Such a system may be desirable to stakeholders who would want to distinguish between candidates who score in the 95th percentile and those who score in the 99th percentile.

The GMAT® uses an 800 point scale that roughly corresponds to z-scores from -4 to +4, where every 100 points on the GMAT® scale corresponds to a standard deviation, and the difference between any two scores is ten points.

The process of putting two different exams on the same scale is commonly referred to as linking, and the process of putting two different scores on the same scale is referred to as equating (de Ayala, 2009).

Item response theory assumes that candidates at a lower ability level will be more likely to get an item wrong than candidates at a higher ability level. As a result, the probability of a correct response from candidates can be charted across all ability levels. This charting of probabilities is frequently referred to as the item response function (IRF) or the item characteristic curve (ICC). It is possible that after pre-testing items, the curve does not look as anticipated. These items should be examined to attempt to understand the cause of the anomaly. For more information on item response functions and equation see He (2010) and de Ayala (2009).

Items are reviewed to ensure that they are testing only the domains they were designed to test. This is a statistical process. Although it is possible to design items that simultaneously test multiple domains, this is regarded as a flaw in an item that was not designed to do so. An example of such an item is a math question that requires a candidate to read a statement, create an algebraic formula from that statement, and use that formula to solve for an unknown value. If the intention of the item was to assess the ability of that student to complete all three steps, then the item would be successful. However, if the intention was to assess the ability of the student to solve for an unknown in an algebraic equation, it may not be successful since it also requires translating a written statement into an algebraic statement. The concept of testing multiple domains may be best considered within the context of a latent trait, referred to above.

Items that exhibit different item characteristic curves between subgroups of the population (frequently referred to as focal groups) and the rest of the population are said to exhibit differentially item functioning (DIF). There are numerous ways to detect such items, which will be explored later in this paper.

3.4 Test Specifications

Now that the items have data from actual candidates, some decisions can be made about how the test will be administered.

Decisions will need to be taken on a number of the exams specifications.

- How can we ensure that all candidates see items from the same domains?

- How does the item selection algorithm work?
- What do we assume is a candidate's starting ability?
- How do we decide when the exam will terminate?
- How big should the item bank be?

3.4.1 Balancing Content.

Although some of these questions may have been answered at the feasibility stage of development, it is important to return to these questions now that data is available from real candidates. Test designers can use the real data to run simulations. The results of those simulations will point to answers to the questions above.

When implementing the ASVAB, the test's designers took the decision not to ensure that all candidates see items on the same topic, which is referred to as content balancing. Based on empirical evidence that suggested the approach would work, test designers balanced the content by ensuring the item bank contained an equal proportion of items testing each domain. The benefit of this design was that the items that would result in the most reliable exam would be administered regardless of the content. The fear was that if the test was constrained to specific content areas, the test would not truly be adaptable.

In contrast the GMAT®'s designers chose to ensure that all candidates would see the same content. The trade off is between reliability and expense. The GMAT®'s designers chose to balance the content by fixing content types to ordinal item positions for an exam. Please see figure 3 below. The designers solved the reliability issue by writing more items. Using the AQA GCSE Math example from above the domain of working with numbers and the number system would be the first item on the exam. Fractions, decimals, and percentages would be the second item on the exam. Ratio and proportion would be the third item. The language of Algebra would be the fourth item. Thus, test designers would ensure that all content areas are tested.

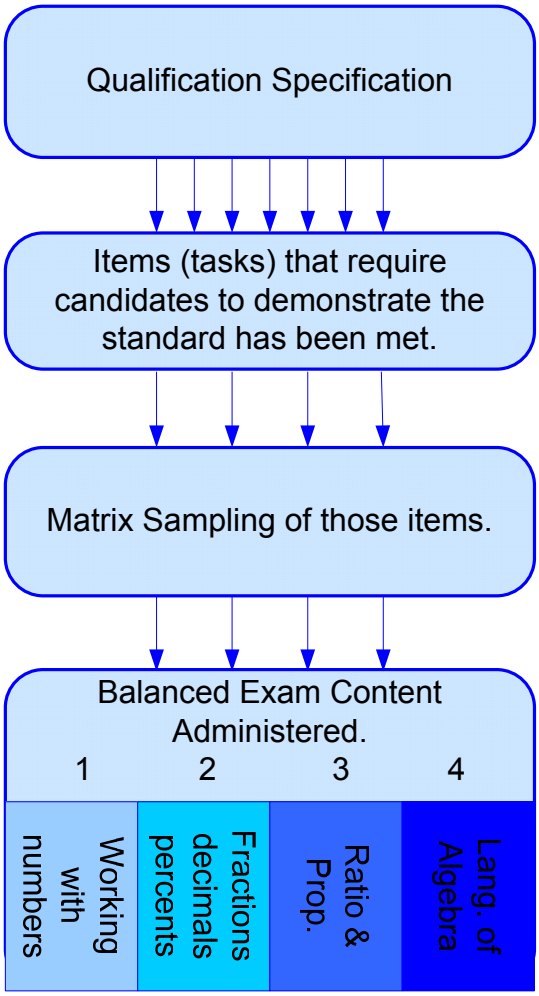


Figure 3.

It should be pointed out that both approaches required the construction of a larger item bank. Thompson *et al.* (2011) suggests that some design specifications of CATs are ultimately based on the lead psychometricians' judgments. Perhaps this is an example of one such specification.

The Regulatory Principles for e-Assessment section 9.4 seems to suggest that the approach taken on the GMAT® would be favored.

3.4.2 The Item Selection Algorithm

The algorithm used for selecting items is often a balance of two somewhat contrasting objectives: the need to select the item that will give the most information about a candidate and the need to ensure that a specific item is not seen so often that

candidates are familiar with it before taking the exam. While these two objectives are not contradictory, in practice the items that will give the most information about a candidate seem to come up very frequently thereby increasing the likelihood that a candidate will have been exposed to them before taking the exam.

The first objective is based on maximizing item information. Information can be thought of as a piece of datum on an item that tells us how confident we are of a candidate's ability level based on the answer given to that item. An exam may be administered with fewer questions if the item selection algorithm always chooses the items with the maximum information. Although it may be useful to think of information in this way, the item's information is in fact a continuous statistical function.

The second objective is based on a common security concern: that candidates will post questions to Facebook, or some other online forum, giving their peers an advantage. To avoid this, the GMAT® has a maximum percentage of items that can overlap from previous pools (Rudner, 2007). Using this approach a candidate would be unlikely to gain a significant advantage by gaining access to the questions from a previous test pool. This is not trivial. The question of exam security is one of the most frequent criticisms of computer based tests, regardless of whether such tests are adaptive (Osterlind and Haverson, 2009)

In practice, an item selection algorithm is chosen that may not choose the items with the most information at the beginning of an exam but will seek items with a greater amount of information toward the end of the exam. These selection algorithms use the number of times an item has been viewed by candidates over its entire life (including the development stage) to demote it in the bank artificially decreasing its information as it is more frequently exposed. Both the GMAT® and the ASVAB use a variation of the Sympson-Hetter exposure control method (Sympson and Hetter, 1985). Controlling for exposure with such an algorithm is in line with sections 9.2 and 10.1 of the Regulatory Principles for e-Assessment.

3.4.3 Determining the Starting Ability

To determine candidate's starting ability level on a computer adaptive test, there are three basic strategies. The first is to start the candidate off in the middle of the difficulty level of the item pool (the mean and the median are both used). The advantage of this strategy is that since there are an equal number of items between either extreme of the item difficulty level, it is likely that the candidate will not exhaust all the items at his or her ability level before a reliable score can be determined.

The second strategy is to start the candidate off with an ability level based on a measure of class performance such as a grade point average (GPA). While GPA will not always be a good indicator of ability level on a standardized test, Thompson *et al.*

(2011) points out that on the whole there will be a gain in efficiency and that a well constructed CAT will compensate if the GPA was not a good starting point.

The third strategy is most applicable to modular testing. This is similar to the GPA strategy. However, rather than using their GPA, candidates' starting abilities are based on the result of the previous CAT administration from previous modules.

3.4.4 Determining the Termination Criteria.

Three strategies can be employed.

The first is based on the standard error of the provisional score. When the standard error is small enough to meet the requirement for the exam, the exam is terminated. The benefit of this strategy is that the fewest possible questions are displayed thus reducing their exposure.

The second is based on a fixed length of the exam. The number of items, the length of time, or a combination of the two can be used. For example, the GMAT® uses both length of time and a fixed number of items. It will end the quantitative section after 37 items have been administered and answered or after 75 minutes, whichever occurs first.

The third strategy is to exhaust the available information in the item bank. The information that items contribute to candidates' scores is cumulative. Items will continue to be administered until the amount of information contributed by each item has fallen below a minimum bound.

The number of items required in the bank will be determined by the level of reliability required for the exam and the other specifications listed above. Ensuring that a item exposure would meet sections 9.1, 9.3, and 12.1 of the Regulatory Principles for e-assessment.

3.5 Publish the Computer Adaptive Test

Thompson *et al.* (2011) points out that this step in test development can be going on concurrently with the other four steps.

Of considerable interest will be the setting of the exam. There are essentially three models of testing that are not mutually exclusive.

- Administration at a purpose built testing center.
- Administration at popup mobile centers, for example in a conference space.
- Administration at schools.

The benefits of a purpose built center include assured standardized quality and security. Invigilators would all be trained employees of the awarding organizations. The

hardware, software, and network connection would be the same ensuring each candidate's experience was largely similar. No operative testing materials would be available outside the testing centers. One of the drawbacks of this approach may be the large investment in real estate. Educational bodies have traditionally invested heavily in real estate. If awarding organizations share this investment strategy, this investment may not be seen as a drawback. If awarding organizations preferred not to invest in real estate, the leases on testing centers would add to the expense of developing a testing program.

Popup mobile centers may convey the same standardized benefits of as purpose built centers. However, moving testing material from place to place could compromise security. Additionally, popup centers prevent computer adaptive tests from truly being on-demand and may prove costly for module testing programs.

Administration at schools could be both cost effective and provide very high security. However, because network connections are not equal in all parts of the country, the delay between exam items could create substantially different testing experiences. Additionally, in this circumstance, invigilators would still need to be mobile traveling from school to school or be available at the schools at all times.

Consultation with awarding organizations will be necessary to take a decision that results in fair exams without adding substantial costs to the operation of a CAT program.

3.5.1 Standardizing the Exam.

The exam conditions will need to be standardized. The standardization should include training proctors, making hardware and software choices, and ensuring that network connections are able to handle the bandwidth of the CAT if it is to be an e-assessment. The standardization will also include the usability of the exam.

Some of the usability standards will be unique to a CAT. For example, in CAT testing candidates can not move backward to review past questions. This creates what is frequently called a mode effect.

Bowels and Pommerich (2001) suggested a method that would permit candidates to review their answers. The concern for whatever bias is introduced by unidirectional linear exams could be alleviated. There is a trade off with this approach. Since the exam is no longer unidirectional, test navigation may result in the introduction of an additional nuisance factor, which could create a bias. Thus one bias would be traded for another.

Wheadon *et al.* (2009) point out that mode effects can limit accessibility to certain question types. For example, on the ASVAB the decision was made not to require scrolling, which shortened reading passages to the length of a computer screen. Interestingly, Pommerich *et al.* (2009) pointed out that when running both a pencil and paper exam and a CAT program, pencil and paper was the limiting mode, not the CAT.

Rudner (2007) mentions item bank rotation and database design and security as two other concerns when publishing an exam. Item bank rotation refers to rotating item pools on a periodic basis. Both the GMAT® and the ASVAB rotate questions on a monthly basis. Thus both exams require candidates to wait one month between retakes so that the question pool is different. While computer adaptive tests are on-demand for first test, they are not technically, on-demand for retakes.

3.5.2 Hardware Considerations

For a web based system, the server side hardware could be cloud hosted by any one of several web companies including Google Aps Hosting, Amazon S3, or Microsoft Cloud. If greater reliability is required of the hardware, test administrators could use mirrored servers, which would require problems with both Google and Amazon before tests would be interrupted. There are several benefits to such a system; the foremost is that these services are scalable and pricing depends on use. In other words, all of the hardware is a variable cost as opposed to a capital investment. While privacy and security is a legitimate concern, hosted servers are already been used successfully to handle private medical data in the United States (Amazon Web Services, 2009).

Using a web based system would also decrease the need for investment in fully functional desktop computers at testing centers. For example, netbooks would be sufficient to take a cloud hosted exam (as long as there were no usability issues). A quick search online reveals netbooks cost as little as £200 each.

While there are database concerns that would need to be addressed, the answer to these concerns is beyond the scope of this paper. Suffice it to say there are concerns which would need to be addressed.

3.5.3 Reporting of Results

Since items are expensive to develop, they cannot be released after use. This can be a problem if the Public do not trust exam results. Releasing some information makes good pedagogical sense, if we consider that assessment is part of the educational process. To balance these needs, it may be wise to construct reports specifically designed for students, teachers, and parents.

Reports should ideally contain as much information as possible about the types of items that the candidates get wrong or right without revealing individual questions. If an item had been tagged when written, the tags could be used as the basis for a report to the student.

Boyle (2010) notes that teachers frequently use past exam papers for students to practise. Since awarding organizations also write curriculums, study question may represent an additional product. Well written practice question books, referred to by reports, would indicate to stakeholders potential areas for improvement.

If results are to be reported, multiple choice items may have advantages over open ended items. Wrong answer choices, frequently called distracters, are often written in anticipation of common student errors; for example forgetting to distribute a negative sign to any term but the first in an algebraic expression. If items were written, banked, and tagged for both the content type and the common mistake used to generate the

wrong answer choice, a report generated after the exam could point to areas of improvement. Wheadon *et al.* (2009) states that, "The case for pedagogical gains from diagnostic information is . . . weak." Even if that turns out to be true, reporting may prove worthwhile both as a communication tool and by allowing the collection of data to further study the pedagogical benefits of reporting. Finally, pedagogical decisions are perhaps best left with educators, as including these stakeholders in the process is more likely to result in a successful outcome.

Another opportunity to improve communication among stakeholders is the release of practice exams. The ASVAB offers six online practice tests that candidates use to decide when they have adequately prepared. A practice GMAT® is also available for free download when candidates register for the exam. One drawback to the GMAT download model is that the software is only compatible with PCs.

Wheadon *et al.* (2009) suggests that for an exam to be reliable, candidates' scores should not increase from repeated sitting of the exam. Perhaps the author meant that scores should not increase from *only retaking* the exam. Otherwise the assumption is that candidates should not be able to raise their score by learning the material being tested.

3.6 Maintaining a CAT

Computer adaptive tests require upkeep to ensure that they remain reliable and valid. Maintenance of the CAT is required to comply with the Regulatory Principles for e-Assessments section 5.2.

3.6.1 Maintaining Reliability

Items can be retired from the item bank because they have been compromised, the parameters drift when recalibrated, or the item has simply been exposed too often.

As a result of retiring items, new items need to be written. There are two types of items that can be written. One is an item clone, which is essentially the same item, which will have been changed. For example, if the item was designed to test the ability to add two three digit numbers, the cloned item would contain two different three digit numbers from the retired item. The benefit of a cloned item is that one can expect it to have the same parameters as the original item. The other type of item is entirely new.

Both types will need to go through the pretesting and data analyses stages before being used in an operational exam. However, new items can be seeded in operational exams. In this way there will be a constant stream of new items being written, evaluated, banked, and administered, decreasing the likelihood of item over exposure.

3.6.2 Maintaining Validity

Maintaining an item pool ensures reliability; participation from stakeholders is required to ensure validity. One benefit of a CAT program is that data become accessible. Newton (2007) pointed out that exams are used by 18 different stake holders.

On November 4, 2010 Ofqual released this tender

> "Investigating the relationship between A level results and prior attainment at GCSE
> Contract Reference: OF113
> The regulators (Ofqual, CCEA and DCELLS) require a technical evaluation of the relationship between A level results and prior attainment at GCSE, and in particular a consideration of whether it can be made more reliable, without becoming too complex or demanding on resources. In addition to this evaluation, the regulators want to investigate the use of this approach to analyse the stability of general qualification results over time."

If GCSE and A levels were administered in an IRT CAT based system, the predictive validity of the GCSE for A level results could be available and updated on a daily basis.

This is meant to serve as an example of how access to data could benefit one stakeholder, but these reports could also be generated for universities, employers, or any of the other stakeholders outlined by Newton (2007).

In addition to meeting the requirement in Regulatory Principles for e-Assessment section 5.4, this practice could actually lower the need for testing. Suppose an employer finds that GCSE results are just as successful a predictor of employee fit as A-level results. The employer may actively seek more employees who have completed a GCSE but not an A-Level. Data would be unique to the employers. So while one employer might find a Math GCSE is a good predictor of success, another employer might find a Business Studies GCSE better for it.

Reports predicting the success of candidates in employment might represent another revenue generating service that awarding organizations could offer. Both the diagnostic and validity reports should be developed with stakeholders while developing the exam to gain the greatest buy in.

If the pilot program is successful, awarding organizations should be prepared to move to a CAT based testing program very rapidly. Pommerich et al. (2009) reports that equating and linking CAT exams with pencil and paper while developing new CAT item pools resulted in "multiple challenges" and an increase in the equating error.

4. Creating an Unbiased Test.

All test developers share a need to ensure that a candidate's score reflects the ability of the domain being tested. When testing population subgroups, this is both a moral imperative and a legal necessity.

Computer adaptive testing offers only two advantages over pencil and paper testing.

- It increases the statistical accuracy of a test, and
- it customizes an exam.

However, these two advantages result in a cascade of potential advantages that can be realized, including

- mode advantages computers offer,
- on demand testing,
- modular assessment,
- eliminating the need for some accommodations,
- decreased testing time,
- improved exam security,
- increased communication to stake holders,
- decreased costs,
- the ability to motivate students,
- improved analysis of item biases, and
- removal of tasks that discriminate against populations subgroups,

While each of these advantages could be realized through other testing models, the greatest potential to realize all of these benefits lies solely with computer adaptive testing.

4.1 Mode Advantages of Computer Adaptive Testing.

Computer based testing has many advantages over pencil and paper testing. Consider how the following scenario, as described in Access Arrangements, Reasonable Adjustments and Special Consideration (JCQ, 2010), might impact a candidate with a hand tremor.

> "A candidate wants to follow an Art course but cannot perform any practical skills independently. The centre requests permission to use a practical assistant. This is refused. It is realised that there are other skills required by the specification which he also cannot fulfil and therefore he decides to follow the course for his education but does not enter for the examinations."

Through the use of traditional media, paper and pencil, the candidate above could not enter examinations for an Art course. The candidate would not be given an equal opportunity to achieve. However, it may be possible to test this candidate on the theory of composition using colored shapes on a computer screen that must be

assembled to meet a design objective. The candidate may not be able to hold a pencil, but may be able to move a cursor on a computer screen either with a mouse or a specially designed pointing device. As an example of technology capable of replacing a mouse, the reader is referred to the Kinect controller for Xbox, which responds to voice commands and body movements.

Further consider that video and audio could be used as part of an assessment. A deaf candidate taking a foreign language exam, for example, who normally reads lips, could have access to video of a speaker, obviating the need for a human reader. Such a video system would lessen the demands placed on test administrators and, as a result, lower costs throughout the testing system.

These benefits may be obtained from a computer based test. However, Osterlind et al., (2009) points out that a computer based assessment is "fundamentally changed by this delivery mechanism." The authors go on to point out that computer based tests are "fraught with psychometric challenges" and that "the mode itself may be a source of measurement error". To overcome these psychometric challenges it is necessary use a statistically robust testing program. The improved statistical accuracy of computer adaptive testing, supported by item response theory, provides the statistical framework necessary to overcome these challenges and minimize the mode impact of computer based tests on population subgroups.

4.2 On-Demand and Modular Testing

If a CAT is made available on-demand and enables modular testing there is an opportunity to further mitigate biases that may be caused by candidate fatigue. A non-native English speaker is likely to become fatigued faster taking an exam since they are applying both English language skills and domain specific skills. A modular testing program could shorten the length of the exam. The results of such a shorter exam would be less likely to disadvantage one group.

4.3 Eliminating the Need for Certain Accommodations

Because CATs can be administered on-demand, they create the possibility to eliminate the need for some accommodations. For example, if a candidate has broken his or her hand but will recover within a week or two of the exam date, the test could be administered when the candidate's hand heals.

4.4 Decreased testing time.

By choosing a statistical ending criterion such as the confidence interval of the score, a CAT can be a variable length exam that continues to ask questions only until the point where enough information exists to confidently assess the candidate's ability. In the United States, The National Council of State Boards of Nursing administers the NCLEX®

uses this termination technique. Items are administered until the system has obtained a 95% confidence interval on the candidates score (The National Council of State Boards of Nursing, undated). The benefit of this system in terms of accessibility is to candidates for whom extended testing periods result in decreased exam performance. For example, an ESL student taking a reading exam may become fatigued more quickly than would a native English speaker. By terminating the exam once enough information has been obtained, the ESL speaker would be less likely to become fatigued. This advantage can be compounded when combined with modular testing program as described above.

4.5 Improved Exam Security.

Since the process of developing a CAT includes the frequent review of items for parameter drift, if cheating were to occur on a CAT it would likely be quickly noticed and the compromised items removed. Further, because exposure control is often accounted for in the item selection algorithm, the impact of one compromised item would not likely be severe.

The adaptive nature of the exam permits for truly unique control mechanisms. For example, the postcode of item writers and school could be part of the data kept on each item. The selection algorithm would then not select items to be administered in the same neighborhoods as they had been written. If item writers work as teachers in the same area as the exam is administered, this precaution would prevent that teacher's students from having an advantage on the exam.

It should be pointed out, however, that if precautions are not taken and exams are run with too small item banks, exam security can be compromised quicker on a CAT then it could be on a paper and pencil test. Since a CAT is more efficient and delivers fewer items, a candidate can remember a greater proportion of exam items from a CAT than the same candidate could remember from a pencil and paper exam. This is compensated for by item exposure controls and large banks of quality items.

4.6 Increased Communication

Assessment, when used to support learning, works best when feedback is given quickly after a task has been completed. Immediate feedback allows students to learn from their mistakes while they can still recall the thinking used on the exam. Since the length of time between sitting an exam and getting the results on current paper and pencil exams is months, students miss the opportunity to learn from their mistakes. Additionally, students are unable to plan retakes before getting results, which can waste a student's time. CATs are scored instantly. As a result of this, students can plan going forward and have an opportunity to consider mistakes they may have made.

Instant scoring is an advantage in terms of motivation, particularly with students who struggle with learning. If a student is given frequent reports outlining improvements, the student's motivation will increase. Like many of the other advantages gained by CATs this type of virtuous feedback is not exclusive to a CAT, but CATs are likely the best way of delivering these advantages.

Further, the exam specification can be tweaked to make testing a little less intimidating. Most frequently a CAT adjusts to a candidate's current ability, attempting to deliver questions at that ability level. As the candidate gets more questions correct the questions get more difficult. There is no reason why a CAT could not be run where at the beginning of an exam candidates are delivered questions slightly below their estimated skill level. Toward the end of the exam, the questions could adjust to be at the candidate's skills level. The effect of using such an approach would be to make the exam feel easier to the candidate, while still gaining enough data to be confident of the candidates score. This advantage is only possible with a computer adaptive test.

4.7 Decreased costs

While initially starting a CAT program is expensive, over time a CAT is less expensive to run than paper and pencil test. As compared to the current testing system, these cost advantages come from decreased testing times, the ability to retain items resulting in the eventual need to employ fewer item writers, and automated scoring resulting in the need not to employ human scorers.

4.8 Improved analysis of item bias

Using an IRT based CAT, permits test developers to use multiple techniques to investigate causes of biased items and tests and to compensate for biased items and tests.

4.8.1 Differential Test and Item Function in Computer Adaptive Testing.

Throughout the development of IRT based CATs there have been a number of terms used to refer to items or tests that disadvantage one group or another. Zumbo (2007) points out that the term *item bias* is imprecise because it has a different meaning to statisticians and non-statisticians. A number of sources point out that an item or test that disadvantages one group can be referred to both by the terms *item/test impact* and *differential item/test functioning* or DIF (Zumbo, 2007 and Bergstrom, Gershon, and Brown, 1993).

Item or test impact refers to an item or test that gives one group an advantage or disadvantage over another group. In the case of item impact, there may be reasons why one group could be at a lower ability level in the latent trait being tested for legitimate reasons. For example, suppose a group of 8 year old students were tested in

a math assessment designed for a group of 12 year old students. It would be reasonable to expect the 8 year old group to perform more poorly than the 12 year old group. In this example, an item that impacted the 8 year old group would not necessarily be disadvantaging 8 year olds. The item would be performing as expected.

Differential item functioning (DIF) refers to an item that gives an advantage or disadvantage to one group over another when the two groups have previously been determined to have the same ability. For example, consider two groups of candidates, one from a comprehensive school and one from a private school, who have previously been found to have the same ability. If the group from the comprehensive school outperformed the students from the private school on a specific item, the item would be displaying DIF.

In practice, test developers may choose to eliminate or compensate for item impact, DIF, and/or differential test functioning.

For example, the GMAT's developers chose to compensate for item impact (Rudner, 2007). The ASVAB's developers chose not to compensate for either because research demonstrated that the outcome of impact and DIF did not change as a result of these items. In other words, the candidates still qualified for the same jobs as they would have had there been no item impact or DIF, so there was no test impact. The ASVAB's administrators note that this is unusual. (Pommerich *et al.*, 2009)

4.8.2 Detecting DIF

DIF is identified as a result of item pretesting. There are numerous statistical methods that can be used for detecting DIF and numerous software packages that can be employed. In general, these methods work by comparing item responses across the ability range between the population and subgroups, between a reference group and a subgroup, or among subgroups.

In general, there are two types of DIF: items that are more difficult for one group than they are for another and items that have a different distribution along the ability range for one group than they have for another (de Ayala, 2009 and Zumbo, 2007).

Test Developers can choose to look differently at the causes of these two types of DIF. One view is that DIF is the result of nuisance factors - outside variables that influence candidates' responses to items. Another view is that DIF can be caused by secondary skills, which are not of primary interest when asking the question but are still of interest. The second view lends itself to an IRT model that views items as testing many skills at once (Zumbo, 2007). This IRT model is frequently referred to as multidimensional item response theory.

The investigation of DIF will follow different routes depending on the view held by the test developer. There are several methods available for detecting DIF including the Mantel-Haenszel procedure, SIBTEST, IRT based DIF methods, and logistics regression methods. While no method is best, an IRT based CAT allows the use of all methods.

The Mantel-Haenszel procedure is perhaps the most frequently used method and compares the odds of a correct response by the focal group to that of the reference group at each ability level tested. While this approach is useful for detecting DIF, it is sometimes difficult to understand the causes of DIF.

Stout, Bolt, Froelich, Habing, Hartz, and Roussos (2003) investigated an approach taken on the Graduate Record Exam (GRE®) reporting that the approach is suitable for the exam. The method, simultaneous item bias test (SIBTEST), requires the analysis of groups, or bundles, of questions at once to determine what are the secondary dimensions (Shealy and Stout, 1993). This multidimensional view of DIF leads to a practice for analyzing causes of DIF. Subject matter experts and psychometricians produce a rubric that represents possible causes of DIF. A SIBTEST is run investigating each of those potential causes. If one of those causes is identified, the item is sent to the item writer with feedback on the potential cause of DIF. The item writer rewrites the item, and the item is tested again. The Ofqual document, Fair Access by Design (Ofqual, 2010), is a good place to start when constructing the rubric. The advantage of the SIBTEST approach is that psychometricians can try to understand the causes of DIF.

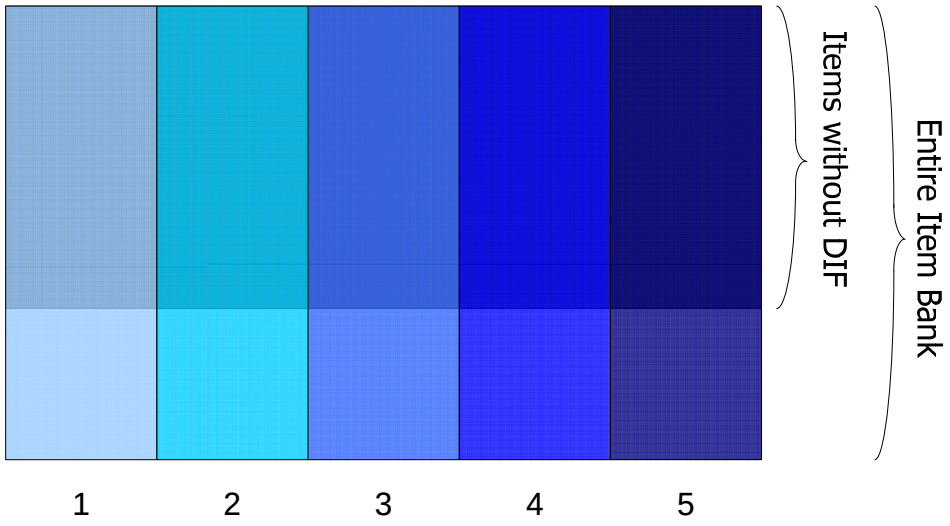
The IRT based approaches of detecting DIF compare the Item Response Function (IRF) for members of a focal group to those of a reference group, or the IRFs for two different focal groups. The advantage of this method is that it permits test developers to understand what specific advantages are gained. The three parameters of an IRT based CAT are the pseudo guessing parameter, the difficulty level, and the ability of the item to differentiate between candidates at specific skill levels. By comparing the two IRFs, test developers are able to determine in what area the focal group is disadvantaged in terms of these three parameters.

The logistics regressions approach is used to investigate conditional dependence between group membership and exam performance.

Because DIF analysis is one of the few areas of IRT based CAT testing that depends on expert judgment, the more information that can be gained to inform those experts the more likely it is that they will make the correct decision. IRT allows for parameter investigation, as a result it gives experts investigating DIF more resources with which to make their judgments.

4.9 Removal of Tasks that Affect a Population Sub-Group

When choosing test specifications, there are numerous methods for eliminating DIF. The first is to simply remove DIF items from an item pool. Zhang, Dorans, and Matthews-López (2005) showed the eliminating DIF items raised the scores of those candidates disadvantaged by those items, while lowering the scores of those candidates who had been given an advantage. However, since there are many subgroups that could be affected by DIF, it may be more economically feasible to eliminate differentially functioning items on a case by case basis. In practice, a teacher or exam administrator could select data about the candidate before the candidate sits the exam. The candidate would be shown only items that are not operating differentially with respect to the candidate's subgroup. These items would not have been specially designed for that subgroup and would have been administered to and calibrated by the entire population. The items administered to the candidate from the subgroup would only be unique in that they would not have displayed impact toward that subgroup. In other words, these items would be equally difficult for the population as they are for the subgroup. It would be possible for candidates who are not in the subgroup to have identical exams to those of candidates in the subgroup. As a result, neither those candidates in the subgroup nor those in the population as a whole would gain any advantage.



Ordinal position of exam items balanced for content.

Figure 4.

This could be accomplished by using the item selection algorithm to choose only from a subset of items. This would be similar to having an expert on accessibility hand picking questions that met the specification but did not adversely affect those candidates in population subgroups.

One way of visualizing this is as Figure 4 above. The entire item bank is represented by all shades of blue. The vertical shades of blue indicate the different content types for the first five items on an exam. The darker regions of those shades at the top represent only those items that display no DIF toward a candidate of a specific subgroup. If the candidate was not a subgroup member the item selection algorithm would be free to choose from any of the items of a specific content type. However, if the candidate were a member of a subgroup, the algorithm would be forced to only choose from the shaded area. This method of compensating for DIF could only be preformed by a computer adaptive test or by a human subject matter expert designing unique test for members of population subgroups. If the subject matter expert did design unique test for members of population subgroups, it is not clear if the results of such an exam would be comparable to the results obtained by the general population.

Another method is to construct testlets during the content balancing portion of the exam. Bao, Dayton, and Hendrikson (2009) showed that since the amount of the advantage is measureable DIF items could be balanced so that while the items may disadvantage one group, the entire exam would not disadvantage any groups. Compensating for DIF in this way was used in the context of reading comprehension exercises, where many questions were asked about the same written passages. Two questions that referred to the same passage may have given an advantage to two opposing subgroups. For example, the first question could have given an advantage to females, and the second question could have given the advantage to males. As long as the magnitudes of the two advantages were similar, both items could be asked, as no one group would experience an advantage on the whole. For example, suppose a candidate was dyslexic. If there was evidence that demonstrated that dyslexic students tend to perform better on spatial items than the average candidate, while performing more poorly than the average candidate on multiplication problems, these two question types would be grouped together. Because the dyslexic student would have an advantage on one content type and a disadvantage on the other, the two questions together would not have display DIF toward a dyslexic student. It is important to point out that this would only be possible if the magnitude of the advantage and disadvantage were comparable. IRT based CAT ensures that test designers have data to make the decision on how to appropriately balance content types. Conceptually, testlets seem well suited to the SIBTEST method of detecting DIF. This method is still relatively new but is promising if the focus is on unbiased tests as opposed to unbiased items.

Rudner (2007) reports that the GMAT® uses a process called equating to compensate for DIF. Recall that equating is the same process that is used to ensure that two different exams are comparable on the same scale. When used to compensate for DIF, equating rescales an item so that it performs equally for the two groups across the ability range. The primary benefits of equating are that it has a proven track record of success and ensures no candidate is disadvantaged.

While we often think of accommodation in terms of a change made to a test that the candidate is aware of, CAT allows for the introduction of accommodations that candidates are unaware of. For example, one of the most frequent accommodations is extra time. Individual items with the same difficulty level can be more time consuming than others. Van der Linden, Scrams, and Schnipke (2003) report using the item selection algorithm as a method to ensure students complete an exam. The reasoning behind the modification is that speed is not the ability being assessed. A simulation was run with less time consuming, though equally difficult, items being delivered to candidates who were performing slowly. Using this approach a CAT would not only adapt to the ability level of candidates, but it would also adapt to the speed of the candidate. While this approach is unlikely to eliminate the need for extra time over one hundred percent, introducing it would allow research to dictate the amount of extra time. As it is now, the amount of extra time offered is arbitrary (Why 25% more time rather than 20%?). Research in the area of timing and an algorithm level solution could ensure that candidates of subgroups receive customized testing experiences. The current research suggests that the combination of the other timing advantages of CATs, on-demand modular tests and shorter exams, with this type of timing accommodation would result in shorter exams that candidates were more likely to finish and that were more statistically accurate and reliable than current paper and pencil tests.

The methods above are possible because a CAT's items have been pretested, thus an administrator has *a priori* knowledge of item functioning. While a paper based test could be constructed using only seeded, pretested items, that test would not present the variety of methods available with CATs for compensating for differential items.

DIF analysis software is available to download for free from a number of different academic websites.

While the methods discussed above relate to how to compensate for DIF once it has been detected, a better method would be to avoid developing items that display DIF. The causes of DIF or impact could be related to decisions made at any stage of test development. With this in mind, this paper will review the steps of test development pointing out opportunities to eliminate DIF.

4.10 Planning to Eliminate DIF (Feasibility).

Test developers must choose between multiple choice and open response questions. Both can affect DIF.

Multiple choice exams may benefit candidates for whom English is not their first language. If these candidates are writing responses to a Science question, but they are not proficient in both reading and writing in English, a multiple choice question could display less DIF, since the candidate would not be incidentally tested in writing while his or her Science skills were being assessed.

Multiple choice questions with tagged wrong responses could be used to help identify candidates from different subgroups. IRT applies the same paradigm that underlies psychological testing. If wrong answer choices were seeded in exams that were especially distracting to members of a subgroup, the exam could detect those candidates. Obviously, those items would only be administered to candidates who have not already been identified as being members of the subgroup. In this way, multiple choice questions would be used for diagnosing not only domain specific skills, but also as a prescreening to diagnose learning disabilities.

Using an electronic reader, such as the one used to grade the GMAT®'s essay section, remains a possibility. Guo (2009) reports that none of the subgroups studied were affected by the use of the IntelliMetric to score essays. While the study is promising, not all of the subgroups outlined in the Ofqual publication `Fair Access by Design' (2010) were studied. The use of an electronic reader would require further study.

During the feasibility stage of development the costs of question rewriting should be considered. The GMAT's developers estimated the cost of creating a question at between 1500 and 2500 USD – a range of 1000 USD. The test developers also estimated that the item bank contained 9000 items. If costs had been toward the lower end of the range, the savings would have been around 9 million USD. While DIF alone is probably not responsible for the entire difference in price, eliminating or rewriting items because of DIF is unlikely to have lowered the price of item development. The GMAT®'s developer's noted that their cost projections initially underestimated the cost of item development (Rudner, 2007). They nearly exhausted the entire cash reserves of the organization that developed the test: the Graduate Management Admissions Council. If awarding organizations want to avoid this unenviable outcome, planning to avoid the rewriting of items due to DIF seems to be sensible.

4.11 Written to be Fair (Item Bank Development).

Just as an item may display DIF, an entire exam may display differential test functioning (DTF) (Bergstom, Gershon, and Brown, 1993). Causes may be unrelated to the items and may include such things as candidates' familiarity with computers, the testing environment, or physical impairments. While there are solutions to these other causes of DIF, all solutions have costs that can, and should, be anticipated.

As pointed out above, item bank development has the potential to be one of the most expensive parts of the transition to CAT based e-assessments. Good item writing practices can result in fewer items displaying DIF. As a result, fewer items will need to be rewritten or removed from exams resulting in cost savings.

Abedi, Leon, Kao, Bayley, Herman and Mundhenk (2011) identify font point size, word length, complex verbs, subordinate clauses, items requiring depth of knowledge, and a

greater percentage of domain specific words in a passage as all contributing to DIF in reading passages administered to eighth graders in the United States.

Everson, Osterlind, Dogan, and Tierre (2007) noted that there is reason to believe that the way a word is referred to in a reading passage is a potential cause of DIF.

Rudner (2007) pointed out that after items are written they should be reviewed for cultural biases.

These findings and practices indicate best practice in item writing. Although some of these practices are undoubtedly already in place in awarding organizations, the opportunity to investigate the impact of these practices prior to administration of an exam, creates the burden of ensuring that no scored exam is administered with a question that is displaying DIF. As such, the cost of a poorly written item will increase. However, if item writing procedures are changed, to allow for feedback to the writers when their items display DIF, the awarding organizations can improve their item writing and reviewing items on the pilot program before making costly mistakes writing items for all subject tests.

For a list of what would be standard considerations when developing items, the reader is referred to Fair Access by Design, Guidance for Qualifications Regulators and Awarding organizations on Designing Inclusive Qualifications (Ofqual, 2010).

4.12 Aiding Teaching (Test Specifications and Publishing)

If the decision was taken, when publishing the exam, to include reporting, the CAT program could enhance teaching. Miller, Chahine, and Childs (2010) report that by using DIF analysis they were able to study teachers and classrooms. Although they concede that further study is required, such a system would provide meaningful feedback to educators as opposed to just exam scores. For example, suppose the students of two teachers of GCSE math have the same average score results. Those teachers could claim to have equal teaching ability. If, however, DIF analysis was used to show that one class performed poorly on simultaneous equation items, while the other class performed poorly on word problem items, it could be shown that both teachers have opportunities to improve their teaching of specific domains of the curriculum.

There is reason to believe that educational programs could learn from ongoing DIF analysis. Using DIF analysis Martinez, Bailey, Kerr, Huang, and Beauregard (2010) demonstrated that withdrawing non-native English speakers from class for special instruction may actually be detrimental to the students. The educational accommodation, English as a second language instruction, could result in students not learning the academic language that would be necessary to perform in their exam. The

intent was to develop a method for detecting missed language opportunities. Because data is so readily available, CAT could be one part of that proposed method.

Jenkins, Levačić, and Vignoles (2006) investigated the effect of school resources on math and science attainment at the GCSE level. The authors point out that the results are inconsistent. Ongoing DIF analysis could result in data for such research being constantly available. Such data could inform school administrators and policy makers who prioritize spending.

5. Conclusion.

5.1 Criticisms of CAT.

Until sufficient technology exists to directly access a candidate's thoughts, no exam is without flaws. CATs are no exception. There are three frequent criticisms of CATs.

- Security can easily be compromised if item banks are small.
- Due to the reduced number of items on an exam, if an item does display DIF and is administered that item will have a greater impact on affected candidates' results.
- Mode considerations.

The first two concerns can be mitigated by developing large item banks and instituting good DIF analysis practices in a test development program. Indeed, much of the recent CAT research has been in the areas of DIF, DTF, and item selection algorithms.

There are mode considerations of a CAT that could disadvantage some candidates. For example, if the hardware is not powerful enough and software is not robust enough there may be a short delay on a CAT between items. For a student who suffers from an attention disorder, this delay could be the source of differential test functioning. If the decision was taken to use electronic readers to grade exams, this would add another second to the delay.

5.2 Recommendations.

Any organization hoping to develop a CAT would benefit by following the same process as followed by the developers of both the GMAT® and the ASVAB.

1. Develop the business case. The benefits of transitioning to CAT should be clear to the awarding organization's board.

    a. Investigation of Revenue Streams.
        i. New Products. New products could include both training products such as online practice tests and preparation questions and training

ii. New Customers. Awarding organizations should consider the possibility that there are untapped markets for testing results. For example, employers or recruitment agencies may be very interested in contacting candidates with high scores in specific subjects if high scores demonstrated predictive validity of job performance for the employer. As long as there was an opt-out option for candidates, such actions would likely be considered beneficial to all parties.

iii. An investigation of liability. Could current practices result in liability for awarding organizations, especially if equal access for all test takers cannot be demonstrated? If CAT testing could eliminate or lessen that liability, the removal of that risk is essentially a increase in revenue.

b. Expenses of Developing a CAT. Based on the product and market investigations, awarding organizations should define and price the development and maintenance of a CAT program.
    i. Personnel Expenses. Awarding organizations will have to survey the internal skills sets within the organization and determine what skills exist within the organization. The likely skills the awarding organizations will need are
        1. Project management skills.
        2. Business development skills. A team to investigate the alternative revenue streams.
        3. Communication skills. Some form of public relations will be necessary to inform the public of the changes to the testing system
        4. Non-mode specific test development skills. These skills most likely exist within awarding organizations.
        5. CAT specific development skills.
            a. Psychometricians.
            b. Statisticians.
            c. Application programmers / developers.

    ii. Delivery expenses. These include leasing agreements, hardware, and software expenses, and reporting expenses. For example, will results be emailed or posted. If posted, the cost for reporting is not insignificant.

iii. Consultation / training expenses. Should some skills be needed on a short term basis, an awarding organization may need to employ a consultant to run the program initially or to train their existing employees.

2. Run a pilot program. Both the GMAT® and the ASVAB ran pilot programs. The results of those programs resulted in increasing the confidence of stakeholders in the program. An awarding organization should run a pilot not only for the benefit of gaining practical experience but also for the purpose of building confidence among stakeholders.

3. Publishing the results of the pilot. The results of the pilot will be of interest to the UK population as a whole. Further the DIF analysis section should be made public to build confidence among those candidates most likely to be compromised by the new testing format.

4. Rolling out the entire CAT testing program. The GMAT®'s developers were able to go through the entire process in three years. Awarding organizations will likely require a similar time frame for each subject test.

Awarding organizations could improve upon these recommendations by offering answers to the following questions:

- Is there a genuine interest in developing CAT programs?
- What research has been done in light of this interest?
- What financial barriers do awarding organizations anticipate to CAT development?
- What technical barriers do awarding organizations anticipate to CAT development?
- What logistical, test delivery, barriers do awarding organizations anticipate? For example would fixed testing centers, mobile testing, in school testing or some other testing center arrangement be preferable.
- What communication problems do awarding organizations anticipate? Communicating the change internally, to employees, and externally, to the public at large, is likely affect the success of any changes.

Answers to these questions will ensure future recommendations are aligned with the goals of the awarding organizations.

A computer adaptive test implemented, using the design outlined above, has benefits that are both numerous and shared among stakeholders. As differential items are found, the reason for those items displaying DIF is used to improve educational experiences through better item writing and improved classroom practices. Priorities can be passed

from a national to school level based on data. Students are constantly challenged at their individual ability levels. Employers make hiring decisions based only on relevant data. Universities can predict the likelihood of an applicant's success based on test results. Awarding organizations could enhance their viability by offering new educational products.

Whilst many testing programs have been switched from pen and paper to computer adaptive testing, there is no evidence of any testing program that has switched back.

6. References.

Abedi, J., Leon, S., Kao, J., Bayley, R., Ewers, N., Herman, J., & Mundhenk, K. (2010). *Accessible reading assessments for students with disabilities: The role of cognitive, grammatical, lexical, and textual/visual features*. (CRESST Report 784). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Available online: http://www.cse.ucla.edu/products/reports/R785.pdf

Amazon Web Services (2009). Creating HIPPA+compliant Medical Data Applications with AWS. Retrieved on February 20, 2011 from http://awsmedia.s3.amazonaws.com/AWS_HIPAA_Whitepaper_Final.pdf

The Assessment and Qualifications Alliance (AQA), (2009). *GCSE Mathematics 2010 Specification (version 2.2.).* Manchester, UK: AQA

Baker, F. B. and Kim, S. (2004). *Item Response Theory Parameter Estimation Techniques, 2nd Edition, Revised and Expanded*. New York, NY: CRC Press, Taylor and Francis Group.

Bao, Han, Dayton, C. Mitchell, & Hendrickson, Amy B. (2009). Differential Item Functioning Amplification and Cancellation in a Reading Test. *Practical Assessment, Research & Evaluation*, 14(19). Available online: http://pareonline.net/getvn.asp?v=14&n=19.

Bergstorm, B. A., Gershon, R.C., and Brown, W. L. (1993) Differential Item Functioning vs. Differential Test Functioning. *Paper Presented at the Annual Meeting of the American Educational Research Association* (Atlanta, GA. April 12 -16) Retrieved on February 20, 2011 from http://www.eric.ed.gov/PDFS/ED377227.pdf

Bowles, R. and Pommerich, M. (2001). An Examination of Item Review on a CAT Using the Specific Information Item Selection Algorithm. *Paper presented at the Annual Meeting of the National Council on Measurement in Education,* Seattle, WA.

Boyle, A., (2010) *Regulatory Research into On-Demand Testing.* Ofqual, Coventry. Retrieved on February 10, 2011 from http://www.ofqual.gov.uk/files/Ofqual-10-4725-Regulatory-research-into-on-demand_testing-2010-03-08.pdf

Cawthon, Stephanie W., Ho, Eching, Patel, Puja G., Potvin, Deborah C., and Trundt, Katherine M. (2009). Multiple Constructs and the Effects of Accommodations on Standardized Test Scores for Students with Disabilities. *Practical Assessment, Research & Evaluation*, 14(18). Available online: http://pareonline.net/getvn.asp?v=14&n=18.

Childs, Ruth A. & Andrew P. Jaciw (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation*, 8(16). Retrieved February 22, 2011 from http://PAREonline.net/getvn.asp?v=8&n=16

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guildfor Press.

Everson, Howard T., Steven J. Osterlind, Enis Dogan and William Tirre (2007). The Performance Effects of Word Locator Cues on the NAEP Reading Assessment. *Practical Assessment Research & Evaluation*, *12*(13). Available online: http://pareonline.net/getvn.asp?v=12&n=13

Gafni, N., Cohen, Y., Roded, K., Baumer, M., & Moshinsky, A. (2009). Applications of CAT in admissions to higher education in Israel: Twenty-two years of experience. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved April 17, 2011 from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/cat09gafni.pdf

Guo, Fanmin (2009). Fairness of Automated Essay Scoring of GMAT® AWA. *GMAC® Research Reports RR-09-01.* Available from http://www.gmac.com/NR/rdonlyres/FACE0811-B6F7-45A9-B57D-ED3703984B9A/0/RR0901_AWAFairness.pdf

He, Q. (2010) Maintaining Standards in on Demand Testing Using Item Response Theory. Ofqual, Coventry. http://e-assessment.org.uk/images/uploads/s-docs/Ofqual-10-4724-Maintaining-standards.pdf Retrieved on February 10, 2011

Jenkins, A., Levačić, R., and Vignoles, A. (2006). Estimating the Relationship between School Resources and Pupil Attainment at GCSE. *Research Report RR727.* Department for Education and Skills, Institute for Education. Retrieved on February 16, 2011 from http://eprints.ioe.ac.uk/1775/1/Jenkins2006Estimating.pdf

Joint Council for Qualifications (2010). Access Arrangements, Reasonable Adjustments, and Special Consideration. London, UK: JCQ. Retrieved on February 25, 2011 from http://www.jcq.org.uk/attachments/published/538/25.%20AARASC%200910.pdf

Martínez, J. F., Bailey, A. L., Kerr, D., Huang, B. H., & Beauregard, S. (2009). *Measuring opportunity to learn and academic language exposure for English language learners in elementary science classrooms.* (CRESST Report 767). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved on February 15, 2011 from http://www.cse.ucla.edu/products/reports/R767.pdf

Miller, Tess, Chahine, Saad & Childs, Ruth A. (2010). Detecting Differential Item Functioning and Differential Step Functioning Due to Differences that Should Matter. *Practical Assessment, Research & Evaluation,* 15(10). Available online: http://pareonline.net/getvn.asp?v=15&n=10.

National Council of State Boards of Nursing (Undated). Computer Adaptive Testing (CAT) Overview. Chicago: National Council of State Boards of Nursing. Retrieved on April 7, 2011 from https://www.ncsbn.org/CAT_Overview.pdf

Newton, Paul E. (2007) 'Clarifying the purposes of educational assessment', *Assessment in Education: Principles, Policy & Practice*, 14:2, 149 -170. Retrieved February 20, 2011 from http://dx.doi.org/10.1080/09695940701478321

Ofqual (2010). Fair Access by Design Guidance for Qualification regulators and awarding organizations on designing inclusive qualifications. *Guidance Document No: 040/2010.* Ofqual, Coventry. Retrieved February 26, 2011 from http://www.rewardinglearning.org.uk/docs/regulation/fair_access_by_design.pdf

Osterlind, Steven J., and Everston, Howard T. (2009) *Differential Item Functioning, Second Edition.* London, UK: SAGE Publications Inc. Series: Quantitative Applications in the Social Sciences.

Pommerich, M., Segall, D.O., & Moreno, K.E. (2009). The nine lives of CAT-ASVAB: Innovations and revelations. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved on February 15, 2011 from www.psych.umn.edu/psylabs/CATCentral/

QCA (2007). Regulatory principles for e-assessment. London, UK: Qualification and Curriculum Authority. Retrieved on February 15, 2011 from http://e-assessment.org.uk/images/uploads/s-docs/Final_regulatory_principles_document_-_PRINTED.pdf

Rudner, L. M. (2007). Implementing the Graduate Management Admission Test® computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved January 10, 2010 from www.psych.umn.edu/psylabs/CATCentral/

Rudner, L. M., Garcia, V., & Welch, C. (2006). An Evaluation of the IntelliMetric SM Essay Scoring System. *Journal of Technology, Learning, and Assessment*, 4(4). Available from http://www.jtla.org

Segall, D. O. and Moreno, K. E. (1999) Development of the CAT-ASVAB. In F. Drasgow & J. B. Olson-Buchanan (Eds.). Innovations in Computerized Assessment (pp. 35—65). Hillsdale, NJ: Lawrence Erlbaum Associates. Retrieved on February 20, 2011 from

http://www.danielsegall.com/catasvab.pdf

Shealy, R., and Stout, W. (1993). A Model-based Standardization Approach that Separates True Bias/DIF from Group Ability Differences and Detects Test Bias/DTF as well as Item Bias/DIF. *Pstchometrika, 58.* 159-194

Stout, W., Bolt, D., Froelich, A., Habing, B., Hartz, S., and Roussos, L. (2003) Development of a SIBTEST Bundle Methodology for Improving Test Equity With Applications for GRE Test Development. *GRE Board Professional Report No. 98-15P, ETS Research Report 03-06*. Princeton, NJ: Educational Testing Service. Retrieved on February 20, 2011 from http://www.ets.org/Media/Research/pdf/RR-03-06-Stout.pdf

Sympson, J. B., & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive tests. *Paper presented at the Annual Conference of the Military Testing Association.* San Diego, CA.

Thompson, Nathan A., & Weiss, David A. (2011). A Framework for the Development of Computerized Adaptive Tests. *Practical Assessment, Research & Evaluation*, 16(1). Available online: http://pareonline.net/getvn.asp?v=16&n=1.

van der Linden, W. J., Scrams, D., and Schnipke, D. L., (2003) Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing. *Law School Admission Council Computerized Testing Report 98-03.* Newtown, PA: Law School Admission Council. Retrieved on February 17, 2011 from http://www.lsas.org/LSACResources/Research/CT/CT-98-03.pdf

van der Linden, W. J. and Glas, A. W. (eds.), (2010) *Elements of Computer Adaptive Testing: Statistics*. Chapters 4, 10, 17, and page 349. London, UK: Springer Science + Business Media LLC.

Wheadon, C., Whitehouse, C., Spalding, V., Tremain, K. and Charman, M. (2009) *Principles and practice of on-demand testing*. Ofqual, Coventry. www.ofqual.gov.uk/files/2009-01-principles-practice-on-demand-testing.pdf

Wise, L. L., Curran, L. T., & McBride, J. R. (1997). CAT-ASVAB Cost and Benefit Analyses. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 227-236). Washington, DC: American Psychological Association.

Wolf, M. K., Kim, J., Kao, J. C., & Rivera, N. M. (2009). *Examining the effectiveness and validity of glossary and read-aloud accommodations for English language learners in a math assessment* (CRESST Report 766). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Retrieved on February 23, 2011 from http://www.cse.ucla.edu/products/reports/R766.pdf

Wortham, J. (2010). With Kinect Controller, Hackers Take Librety. *New York Times Online.* November 21, 2010. Retrieved February 28, 2011 from http://www.nytimes.com/2010/11/22/technology/22hack.html.

Wu, Brad (2010). Technical Report UK Clinical Aptitude Test (UKCAT) Consortium Testing Interval: 7 July 2009 – 10 October 2009 Executive Summary. Retrieved April 17, 2011 from http://www.ukcat.ac.uk/pdf/UKCAT%202009%20Technical%20Report_abridged%20_3_.pdf

Yanling Zhang, Neil J. Dorans, and Joy L. Matthews-López (2005). Using DIF Dissection Method to Assess Effects of Item Deletion. *Research Report No. 2005-10.* The College Board, New York. Retrieved on February 19, 2011 from http://professionals.collegeboard.com/profdownload/pdf/051766CBReport2005-10WEB.pdf

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where it Has Been, Where it is Now, and Where it is Going. *Language Assessment Quarterly*, 4(2), 223-233, Lawrence Erlbaum Associates, Publishers. Retrieved February 20, 2011 from http://educ.ubc.ca/faculty/zumbo/papers/Zumbo_LAQ_reprint.pdf

7. Acknowledgements

The author wishes to thank Lawrence Rudner of the Graduate Management Admission Council, Mary Pommerich of the Defense Manpower Data Center, and Paul Edelblut of Vantage Learning, who all suggested research in support of this document.

First published by the Office of Qualifications and Examinations Regulation in 2011.

Office of Qualifications and Examinations Regulation Spring Place Herald Avenue
Coventry Business Park Coventry CV5 6UB