Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 16, Number 1, January 2011 ISSN 1531-7714

A Framework for the Development of Computerized Adaptive Tests

Nathan A. Thompson, *Assessment Systems Corporation*
David J. Weiss, *University of Minnesota*

A substantial amount of research has been conducted over the past 40 years on technical aspects of computerized adaptive testing (CAT), such as item selection algorithms, item exposure controls, and termination criteria. However, there is little literature providing practical guidance on the development of a CAT. This paper seeks to collate some of the available research methodologies into a general framework for the development of any CAT assessment.

Computerized adaptive testing (CAT) is a sophisticated method of delivering examinations, and has nearly 40 years of technical research supporting it. An additional body of literature investigates the context of CAT, such as comparisons to paper-based or computer-administered conventional tests (Vispoel, Rocklin, & Wang, 1994) and the application of the CAT approach to specific tests (Sands, Waters, & McBride, 1997; Gibbons et al., 2008). However, except for some coverage within technical books such as Flaugher's (2000) discussion of item banks or discussions of practical issues such as Wise and Kingsbury (2000) or Parshall, Spray, Kalohn, and Davey (2006), little attention has been given to the test development process in the CAT context. Moreover, research and recommendations have not been consolidated to produce a general model for CAT development. The purpose of this paper is to present such a model for the development of a CAT assessment program, which is general enough to be relevant to all assessment programs but specific enough to provide guidance to those new to CAT. A particular focus is given to the necessity of simulation research to adequately answer questions encountered during the development of a CAT.

The framework (Table 1) is intended to cover the entire process of CAT development, from inception to publication rather than just psychometric aspects. Therefore, it begins not with the decision to implement CAT, but rather when the question is raised as to whether CAT might even be an appropriate test administration method for a given assessment program. Several important questions need to be answered before the development of an item bank or delivery platform. Only then can the test development process proceed with the steps shown in Table 1.

Table 1: Proposed CAT framework

Step	Stage	Primary work
1	Feasibility, applicability, and planning studies	Monte Carlo simulation; business case evaluation
2	Develop item bank content or utilize existing bank	Item writing and review
3	Pretest and calibrate item bank	Pretesting; item analysis
4	Determine specifications for final CAT	Post-hoc or hybrid simulations
5	Publish live CAT	Publishing and distribution; software development

This paper proceeds to discuss some of the issues relevant to each stage. This discussion, however, is by no means comprehensive. To the extent that each assessment program's situation is different and unique, it raises its own issues. Moreover, extensive attention has been given to individual aspects in other sources, such as the technical discussion of item exposure in Georgiadou, Triantafillou, and Economides (2007). Therefore, an assessment program should utilize this framework as simply that, rather than as a comprehensive recipe, to identify issues relevant to the situation at hand and the type of research, business, or psychometric work necessary to present guidance for each decision.

This is important not only from a practical viewpoint, but because this is the foundation for validity. A CAT developed without adequate research and documentation in each of these stages runs the danger of being inefficient at the least and legally indefensible at the worst. For example, arbitrarily setting specifications for a live CAT (termination criterion, maximum items, etc.) without empirical evidence for the choices could result in examinee scores that are simply not as accurate as claimed, providing some subtraction from the validity of their interpretations.

Background

While the details regarding CAT as a delivery algorithm are discussed at length in numerous sources (e.g. Lord, 1980; Wainer, 2000, van der Linden and Glass, 2010), some background is necessary to provide a frame of reference for discussions.

From an architectural perspective, a CAT is composed of five components (Weiss & Kingsbury, 1984; Thompson, 2007). The first component is a calibrated item bank, and is therefore developed as test content (e.g., mathematics items for a mathematics exam). The remaining four components are psychometric rather than content, and refer to algorithms in the CAT system.

1. Calibrated item bank
2. Starting point
3. Item selection algorithm
4. Scoring algorithm
5. Termination criterion.

A CAT operates by taking the first two components as a given, then cycling through 3, 4, and 5 until the termination criterion is satisfied (Figure 1). For example, an examinee sits at a computer to take a test. The computer is preloaded with the item bank (which includes psychometric data on each item), and a specific starting point will have been determined for the examinee. An item is selected for this starting point, the first item in the test. After the item is answered, it will be scored and an estimate of examinee ability (θ) obtained. The termination criterion will then be evaluated; if it is not yet satisfied, another item will be selected (component 3), which the examinee will answer, then the examinee's score (θ) is updated (component 4), and the termination criterion evaluated once more (component 5).

Because the delivery of a CAT is a collaboration between these algorithms, it is just as important to establish appropriate specifications for the algorithms as it is to develop an appropriate item bank. This process of research to determine specifications is not widely understood, and is typically left purely to the professional opinion of the psychometrician in charge of the testing program. This paper not only provides a model for this process that psychometricians can follow, but also elucidate some of the issues for non-psychometricians who are nevertheless stakeholders in the process and responsible for some of the work required in the process.

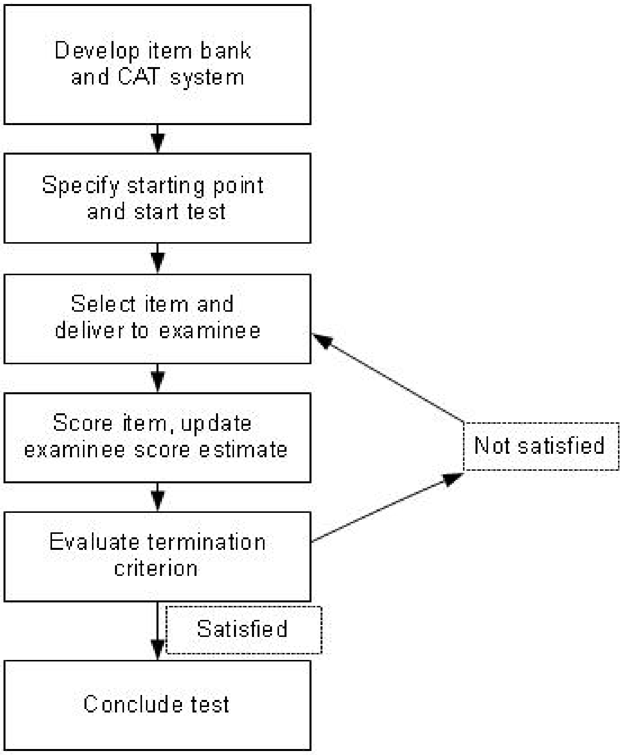
Develop item bank and CAT system
Specify starting point and start test
Select item and deliver to examinee
Score item, update examinee score estimate
Evaluate termination criterion
Not satisfied
Satisfied
Conclude test

Figure 1: Example flowchart of CAT algorithm

Most CATs are constructed on the foundation of item response theory (IRT). IRT is a powerful psychometric paradigm with many advantages for test development, item analysis, and scoring of examinees. With regard to CAT, the most important advantage is that it places items and examinees on the same scale, facilitating the direct matching of examinees to items that are most appropriate for them. While CATs can still be designed with classical test theory (Frick, 1992; Rudner, 2002; Rudner & Guo, in press), this advantage means that the vast majority of CATS are based on IRT. Therefore, a level of familiarity with IRT is necessary to understand CAT. The uninitiated reader is referred to Embretson and Reise (2000) or de Ayala (2009). While an effort is made to provide as broad and general a framework as possible, the perspective of this paper is primarily limited to CATs based on IRT because of its advantages and prevalence in the field. The framework would need to be adapted somewhat for CATs based on classical test theory, or tests that are not fully adaptive, such as fixed programmed branching or multistage testing, but the principles remain applicable.

Step 1: Feasibility, applicability, and planning studies

The first stage in CAT development is to determine whether the CAT approach is even feasible for a testing program. Because the CAT algorithm is so conceptually appealing and offers certain well-known advantages, non-psychometrician stakeholders might become enamored of the idea and wish to proceed without knowing anything about CAT. An executive or professor might hear that CAT typically uses only half as many items as a conventional test (Weiss & Kingsbury, 1984) or even less, and simply make a decision that the testing program will move to CAT. This can be quite dangerous, not only from a psychometric point of view, but also from a business perspective. Transforming an assessment program from fixed-form tests to CAT is not a decision to be made lightly.

Therefore, the practical and business considerations should be researched first. Does the organization have the psychometric expertise, or is it able to afford it if an external consultant is used? Does the organization have the capacity to develop extensive item banks? Is an affordable CAT delivery engine available for use, or does the organization have the resources to develop its own? Will converting the test to CAT likely bring the expected reduction in test length? Does the reduction in test length translate to enough saved examinee seat time – which can be costly – to translate into actual monetary savings? Or even if CAT costs more and does not substantially decrease seat time, is that fact sufficiently offset by the increase in precision and security to make it worthwhile for the organization?

Fortunately, many such questions can be answered not simply by conjecture, but by psychometric research. Monte Carlo simulation studies (van der Linden & Glas, 2010) can allow a researcher to estimate not only the test length and score precision that CAT would produce, but also to evaluate issues such as item exposure and the size of item bank necessary to produce the desired precision of examinee scores. These studies operate by simulating CATs under varying conditions for a large number of imaginary examinees. The results can then be compared to make decisions. For example, CATs could be simulated for a bank of 300 items and a bank of 500 items, and results compared to determine which presents a better goal for the organization. What makes this approach so important at this stage is that Monte Carlo studies can be done before a single item is written or before any real data is available.

Monte Carlo simulations are based on the fact that IRT provides an estimate of the exact probability of a correct response to an item for a given value of θ. This allows researchers to easily generate a response to an item, given its item parameters and a value of θ. For example, supposed that an average examinee ($\theta = 0.0$) is calculated to have a 0.75 probability of a correct response to an item. A random number is generated from a uniform distribution with a range of 0 to 1. If the value is 0.75 or less, the generated response is "correct." If the value is greater than 0.75, then the generated response is "incorrect." Given item parameters for a bank and a sample of examinee θ values, an entire dataset of correct/incorrect responses can be easily generated. The item and examinee parameters can be real or randomly generated themselves, depending on the availability of data at a given stage of the CAT development process. If randomly generated, basing the generation on expected parameters makes the simulation more defensible. If similar tests in published research have been found to have an average discrimination parameter of 0.7, then it obviously makes sense to generate an item bank that reflects this fact.

This dataset can then be used to simulate CATs. Simulated CATs operate the same as live CATs, with the exception that the item response is not provided by a live examinee, but rather looked up in the table of generated responses or generated in real time. If the CAT selects a certain item to be administered, the simulation program simply provides the response from the data set.

Because Monte Carlo CAT simulations can only be done with specialized software, the first step is to obtain the necessary software. Two pieces of software are necessary: one to generate a data set based on specifications you provide, and one to simulate how CAT would perform. WINGEN (Han, 2007) and PARDSIM (Yoes, 1997) can simulate data sets based on item response theory (IRT; Embretson & Reise, 2000) under a wide range of specifications. CAT tests can then be simulated using FireStar (Choi, 2009) or CATSim (Weiss & Guyer, 2010). CATSim advantageously combines the two pieces, and can simulate its own Monte Carlo data sets, utilize real data sets, or perform a hybrid of the two, in concert with CAT simulation. Alternatively, if a testing program has substantial psychometric expertise, simulation software can be developed in-house, but the cost in hours will most likely exceed the cost of obtaining existing software.

There are several important dependent variables to consider in Monte Carlo simulations. The two most important are average test length and the precision of the test, quantified as the standard error of measurement. With conventional tests, the test length is fixed but the precision is variable; examinees in the center of the distribution typically have less error with regards to measuring their latent ability because items of medium difficulty are the most common. With adaptive tests, test length is typically variable, but the CAT is designed to provide equivalent precision for all examinees if the item bank is properly designed, one reason that effective simulations are essential.

The next step in this stage is to make business case evaluations based on the results of the Monte Carlo studies. For example, suppose that a testing program currently utilizes four conventional fixed-form tests of 100 items, with 20 items of overlap for equating. This translates to a bank of 340 items. It might have been initially thought that moving to CAT would require a bank of 1,000 items at the very least, but Monte Carlo simulations showed that a bank of 500 items is adequate. Considering that the bank currently stands at 340 items, the additional item development costs would be much smaller than originally expected. Furthermore, the simulations showed that the bank of 500 items could produce tests that were as precise as the current tests, but with an average of 55 items. Would the cost of developing 160 new items, performing the necessary CAT research, and moving to a CAT testing engine be offset by the time savings of 45 items per examinee and the additional security by using more than four forms? Those are the types of questions that are the crux of this step, but should also take into consideration non-business advantages, such as being able to measure all examinees with equal precision or an ameliorated examinee experience due to seeing only appropriate items.

Step 2: Develop item bank content

Once the final decision has been made to convert to CAT, the next step is to establish an item bank. Again, this should be done based on empirical evidence when possible. The simulation studies in the previous step should be utilized and probably expanded to provide guidelines for the bank; as noted by Veldkamp and van der Linden (2010), simulations are useful for this step and not necessarily limited to use after pilot testing as described in Flaugher (2000). Not only is the number of items in the bank important, but also the distributions of item parameters and practical considerations such as content distribution and anticipated item exposure issues. Simulations should be completed with various situations, such as a bank with a wide range of difficulty compared to a narrow range, or skewed difficulty, or a bank with more highly discriminating items compared to less discriminating items. Veldkamp and van der Linden also discuss optimal bank research; the Reckase (2003) approach can provide valuable information.

An important consideration in designing the studies is that the test information function (TIF; Embretson & Reise, 2000) should match the purposes of the test. If the test is used for classifying examinees based on a single cutscore (e.g., pass/fail), the test requires more information near that cutscore than it does on the extremes of the ability range. Precise scores are not needed for examinees on the extreme, so items of extreme difficulty are not necessary. Conversely, if precise scores are needed for all examinees, including those of very high or low ability, then items appropriate for those examinees are needed. Substantial numbers of very easy or very difficult items are required.

Fortunately, in many cases a completely new item bank is not necessary. The existing item bank can be utilized. In fact, it is often quite useful to do so for continuity purposes. By linking and mixing newly developed items with an existing bank, this ensures that the underlying IRT scale remains constant during the transition to CAT. Of course, doing so also greatly reduces the number of items that need to be developed.

Regardless of whether the bank will consist of all new items or a mix of old and new, it is important to consider the statistical requirements of items in a testing program. If a testing program has high standards and typically eliminates a substantial percentage of items during the development process, this must also be taken into account during this stage.

Step 3: Pretesting, calibrating, and linking

Once items are developed, they must be pretested. This is absolutely essential for CAT because items are matched to examinees based on IRT item parameters, and the parameters are estimated via statistical analysis of actual examinee responses to items. The sample size required for pretesting varies by the IRT model employed (Embretson & Reise, 2000). For example, Yoes (1995) suggests that 500 to 1,000 examinees are needed per item for the three-parameter IRT model. Topics in this step are described in more detail in Flaugher (2000).

There are two approaches to pretesting, referencing the previous issue of whether the CAT item bank will be completely new or a mix of old and new, and whether the existing tests must remain operational during the item development and pretesting phase. If the CAT bank will be completely new, the items can simply be administered in large numbers; in developing a bank of 400 new items, each examinee might have the time to see 100 new items. If there is a mix of old and new, and the current tests must remain operational, the new items might be “seeded” into the currently operational tests. Let us continue with the previous example, where 160 new items were needed in addition to 340 existing items. To account for the fact that some items will not turn out as good as hoped for, suppose we are pretesting 200 items. If examinees are already taking a 100-item fixed-form test, taking all 200 new items would triple the test length, which would take up too much time. Since 200 items are needed, and there are four forms, it makes sense to give only 50 new items to each examinee. The 50 items can be selected randomly, or in predefined blocks using various plans (Verschoor, 2010). The key, regardless, is to plan the arrangement of pretest items such that enough examinees see each item to provide the minimum number of responses needed.

After pretesting is completed, the item parameters must be estimated with IRT calibration software. An important component of this is *linking*, which ensures that parameters from all the items are calibrated on a common scale. There are several approaches for this, but one important distinction needs to be made, between methods that put the new items on an existing scale (e.g., Stocking & Lord, 1980) or methods that establish a new metric (Lee & Weiss, 2010). Obviously, if the item bank is to be completely new, there is rarely a need to link it to an existing scale. Similarly, if the bank is being designed to incorporate items from an existing test and it is necessary to maintain the scale, then a method that establishes a new metric is inappropriate. For guidance on linking, refer to Kolen and Brennan (2004).

This calibration phase involves additional statistical analysis. Most commonly, item statistics such as difficulty and discrimination are reviewed to determine if items need to be eliminated or revised and pretested again. Even if the testing program is officially based on IRT, classical statistics can still be quite useful for this purpose. An additional statistic at the item level is the analysis of *model fit*, namely how well the data supports the IRT model that has been assumed for the calibration. Items that have substantial issues, such as speededness or susceptibility to guessing, will typically have poor fit, which implies that IRT parameters for those items are not stable enough to be used in CAT.

Lastly, an analysis of dimensionality is necessary at this stage. IRT assumes that the test is unidimensional (unless multidimensional IRT models are employed), so the items in the pretesting of the bank should be factor analyzed to ensure this. The appropriate procedure is factor analysis using tetrachoric correlations (for items scored correct/incorrect), which can be done with the software program MicroFACT (Waller, 1997), or *full-information factor analysis* using TESTFACT 4 (Bock et al., 2003). Bejar (1980; 1988) has suggested an alternative method of evaluating dimensionality within the IRT framework.

Step 4: Determine specifications for the final CAT

At this point, an item bank has been developed and calibrated with IRT. However, this is only the first of five components of a CAT described previously. Before the CAT can be published and distributed, the remaining four components must be defined. As with the planning of the item bank, this should not be done based on arbitrary decisions, but on simulation studies (Flaugher, 2000). However, there is one important difference in this stage: we now have an actual item bank developed and data from real examinees responding to those items. Real data is obviously preferable to randomly generated data if the purpose is to approximate how the CAT will perform with real examinees in the future. Therefore, this data can be utilized in new simulation studies, called *post-hoc simulation* or *real-data simulation.*

With post-hoc simulation, like Monte Carlo simulation, a CAT is simulated for each examinee based on responses to each item in the bank. The difference is that Monte Carlo simulation generates the response of each examinee to each item, while post-hoc simulation utilizes the real data. For example, if the CAT simulation for the first examinee determines that Item 19 from the bank should be the first item administered, Monte Carlo simulation would generate a response to that item based on the item parameters, the person parameter (θ), and the assumed IRT model. On the other hand, with post-hoc simulation there would be no need to generate the response; the simulation algorithm would simply look up the actual response of the first examinee to Item 19.

This type of simulation has a substantial drawback with pretest designs where examinees saw a small percentage of the items in the bank. In the example above, each examinee would see only 150 items from the developed bank of 540 (with the intention that 500 would be retained): 100 items from an existing form and 50 new items. If a post-hoc simulation were to be conducted on this data set, a response would not be available for 390 items for each examinee. To address this issue, a third type of simulation, *hybrid simulation*, was developed (Weiss & Nydick, 2009; Weiss & Guyer, 2010). Real data is used where available, but missing responses are generated using Monte Carlo methods based on each examinee's *θ* as estimated from the items he/she has answered. This allows CATs to be simulated more effectively with a real item bank and real examinees.

Post-hoc or hybrid simulations are essential to compare and evaluate different methods and specifications for the four algorithmic components of CAT with a real item bank. There are often important questions to be answered within each component, such as comparing item exposure methods or applying content constraints in the item selection algorithm; software such as CATSim (Weiss & Guyer, 2010) is designed to provide options to specifically answer such questions. A CAT that is published without adequate research in the form of these simulation studies is substantially less defensible. For example, the item bank might be inadequate to meet the demands of the item selection, content balancing or termination criterion algorithms; without simulation studies, this might not be realized until after the tests are in the field.

Item bank

The item bank does not necessarily have to be used as is. While a bank of 500 items has been developed, perhaps the items are higher quality than expected, and a bank of 400 might suffice, allowing the other 100 items to be rotated into position at a later date. Simulations could easily compare CATs with all 500 items to CATs with only 400 items from the bank.

Starting point

There are several options available as the starting θ estimate assigned to each examinee before an item is administered. The most straightforward is simply to assign a fixed value corresponding to an average score. With IRT, this is usually 0.0 because the scale is centered on examinees.

Starting each examinee with the same initial θ estimate has a distinct disadvantage. Because the CAT algorithm selects the best item for an examinee based on the θ estimate, if every examinee has the same estimate, than every examinee will receive the same first item. If this is deemed to be a test security or item exposure issue, some randomization can be implemented. For example, the estimate can be a value randomly selected in the range -0.5 to +0.5, or a randomesque item selection method applied, either of which would likely enable several possible starting items.

Nevertheless, the goal of CAT is to adapt the test to each examinee as much as possible. Both of the previously mentioned starting points assume that nothing is known about the examinee. However, in many cases there is information available on examinees. The most obvious is scores on previous tests. If CATs are being administered to children in schools as part of a formative assessment program, they are often used several times per year. In such a situation, the score from the first administration makes an ideal starting point for later administrations, because student ability will likely be in a similar range, though will hopefully increase to some degree.

Another option is to use external information to estimate examinee ability. For example, Castro, Suarez, and Chirinos (2010) examined external factors like motivation and socioeconomic status. In educational contexts, other assessments or scholastic information can be useful. For example, with a test for professional licensure or certification that is taken after the educational process, performance indicators from the process, such as grade-point average, could be used as a starting point if research shows that there is a correlation. While not a perfect prediction for every examinee, this would provide an increase in efficiency, on average, that could translate to substantial time and item exposure savings in the long term. For the minority of examinees where there is an inaccurate prediction, the adaptive nature of the CAT will account for it.

Item selection algorithm

The item selection algorithm is important because it refers not only to the specific calculations to determine the most appropriate item, but also to the impact of practical constraints. Item selection is typically based on the concept of *item information*, which seeks to quantify the notion that some items are more appropriate than others for a certain situation. For example, it makes little sense to administer a very easy item to an examinee that is quite bright; they are virtually guaranteed to get it correct. The converse is true for an examinee of low ability.

An important consideration in item selection is whether the purpose of the test is to obtain accurate point estimates of θ or to make broad decisions. If the purpose of the test is to estimate θ with a certain level of precision, then it is appropriate to deliver items that provide the most information at the θ estimate of the examinee. However, if the purpose of the test is to classify examinees based on a cutscore, using a likelihood ratio approach (Reckase, 1983), it is often more efficient to design the item selection algorithm to evaluate information at the cutscore (Eggen, 1999; Eggen & Straetmans, 2000; Thompson, 2009).

There are a number of methods of calculating the IRT information criterion used to select items, and a substantial amount of CAT research consists of simulation studies designed to compare different methods of item selection (e.g., Eggen, 1999; Weissman, 2004). The 2010 International Association for Computerized Adaptive Testing conference included two sessions devoted directly to research on item selection algorithms. Yet in practice, these differences are often insignificant; for this reason, it has been argued that other avenues of making the test more efficient should be evaluated (Thompson, 2009; van der Linden, 2010).

For the same reason, it is often more important to evaluate the impact of practical constraints in the item selection process. The two most common types of constraints are item exposure constraints and item characteristic constraints. Item exposure constraints are subalgorithms incorporated into the item selection algorithm to combat the fact that CAT always tries to select the best items, which tend to be the items with the highest discrimination parameter. Therefore, items with higher discrimination parameters are administered far more often than items with moderate or low discrimination. To address this, some type of randomization is typically implemented. See Economides, Georgidou, and Triantfillou (2007) for a review of these methods.

Many testing programs also require that tests be constrained by certain non-psychometric characteristics. A typical example of this is content constraints, such as a math test requiring a certain percentage of items covering algebra, geometry, and probability. Another example is cognitive level, including Bloom's (1956) taxonomy, which might require that no more than a certain percentage of the test be simple recall questions.

Both of these types of constraints reduce the efficiency of the adaptive algorithm because they impede the natural selection process of choosing the most discriminating items. However, they can be quite important from a broader perspective. Therefore, post-hoc or hybrid simulations should take them into account when determining CAT specifications, and provide detailed guidance regarding their use. Not only are the simulations useful for evaluating the application of item exposure constraints, but also for comparing the efficiency of different methods of controlling item exposure.

Scoring algorithm (θ estimation)

Most CATs utilize IRT for scoring, in addition to item selection. Although Rudner (2002) showed that CATs designed with classical test theory can be quite efficient in the classification of examinees, CATs for point estimation of examinee ability require the precision that IRT can provide. Simulation studies can be used to compare the efficiency of CATs designed with different scoring algorithms. This not only includes classical vs. IRT, but also a comparison of IRT methods, such as maximum likelihood and Bayesian methods. The latter comparison produces little difference in observed results, but does have some important implications. Maximum likelihood estimation is less biased (Lord, 1986), but has the drawback that it requires mixed response patterns (at least one correct and one incorrect response), which is never the case after the first item is administered. A subalgorithm must then be applied when there is a nonmixed response vector; simulations can also aid in that specification.

Termination criterion

While CATs can be designed to be fixed length (e.g., all examinees receive 100 items, but the items are adaptively selected from the bank), they enable the possibility of *variable-length* tests. Such a test not only adapts the items to the examinee, but also adapts to the number of items needed. There are different methods to implement this. Some evaluate the examinee θ estimate, some the standard error of measurement, and some take into account the item bank.

An example of a termination criterion based on the θ estimate is to terminate the test when the θ estimate no longer changes more than a small amount after each item. This is because CAT is an iterative process, so the estimate typically varies widely as a test begins, but eventually "zeros in" on examinee ability. The same is true for the standard error of measurement; it is relatively large at the beginning, and will decrease as the test proceeds.

Another approach is to base the termination criterion on the item bank rather than an examinee parameter. One example of this is the *minimum information criterion*; if there are no items left in the bank that provide at least some minimal level of information, as defined by the item selection algorithm, then the test can be stopped because there are no more items left that are worth administering.

However, the most common termination criterion is the *minimum standard error criterion*. This approach designs the test to stop when an examinee has reached a certain standard error, or equivalently, a certain level of precision. For instance, the test might stop when the standard error becomes 0.25 or less. This would mean that a 95% confidence interval with ±2 standard errors on each side would be approximately one θ unit wide. This termination criterion has the advantage of producing *equiprecise* scores for all examinees, assuming that the item bank is properly developed.

Like item selection, this algorithm is also subject to practical constraints. The typical constraint is a test length constraint, in the form of a minimum or maximum. The minimum serves to ensure that each examinee receives at least a certain number of items; if the test can fail examinees with as few as 10 items, then it might be politically advantageous to ensure that examinees see at least 20 items before failing, in an effort to reduce complaints. The maximum serves to ensure that the entire bank is not administered. In a pass/fail CAT, examinees whose true θ is equal to the cutscore will never be able to be definitively classified even if given the entire bank, so the test might be set to terminate at some relatively large number like 200 items.

These options all provide direct control over the operation of the CAT and directly affect the number of items seen by examinees. In general, a test with more items produces more precise scores, and vice-versa. Simulation studies are necessary to evaluate the extent of this tradeoff and produce test specifications that meet the requirements of the testing program. If the minimum standard error criterion is employed, it would be useful to run simulations with varying levels of error, perhaps 0.25, 0.30, and 0.35, then evaluate the greater number of items required for greater precision.

Step 5: Publish live CAT

Once the specifications for all the necessary components have been determined, as well as any additional algorithms, the final CAT can be published. If the test development and delivery software already exists (for example, the organization has purchased a system or access to a system), this step contains little difficulty. Most of the options described in the previous section are manifested as simple radio buttons or check boxes within the CAT system. However, if the organization is developing its own platform, this step can be the most difficult. Fortunately, if that is case, most of the development work can be done concurrently with the previous four steps, saving a substantial amount of time. This step also contains many of the practical distribution and delivery issues and effort that pertains to all testing, not just CAT, such as test security.

Epilogue: maintaining a CAT

The research involved in CAT development does not cease when the test is published. Additional research is needed as maintenance for the CAT. Perhaps the most important thing to check is whether actual CAT results after publication match the results expected based on the simulations. For example, if post-hoc simulations predicted that examinees would need 47 items on average to reach the minimum standard error of 0.25, did this actually occur during the first month of operational CAT?

Another important issue is maintenance of the item bank, sometimes called "refreshing." Because items can become overexposed in large volume testing, overexposed items might need to be rotated out and newer items rotated in. This is typically done by seeding new items into the bank to be pretested, and then converted to scored items after sufficient sample size for calibration is obtained. However, some research has investigated the application of online calibration, where the items are immediately calibrated into the bank during the pretesting process.

The selection of items to be retired is a choice of the test sponsors. There are several issues to consider. The most obvious is exposure; if half the examinees see a certain item, and it is known that items typically find their way to the Internet, then the item can likely be considered compromised. A more specific method of examining this issue is a parameter drift study. If the item is compromised, then many more examinees will answer it correctly than when the item was first developed. If post-compromise data is analyzed, the IRT item parameters will then be different, indicating that the item should be retired. Test security software designed to search the Internet for test items at brain dump sites is also useful.

Summary

The development of a CAT requires substantial psychometric expertise. Because of this, the development of a CAT is often left completely to the judgment of the professionals working on the CAT. But as CATs become more widespread, the psychometric expertise of the personnel working on them might not be sufficient to develop a legally defensible CAT without some guidance. This paper has provided a broad framework for the development of a CAT, applicable to most situations. However, although this model is quite general, and many issues have been discussed, it is not completely comprehensive. Many testing programs will have idiosyncratic issues that must not only be identified, but also isolated so that they can be investigated as empirically as possible. However, the principle that answers to the issues should be empirically identified, often through the use of simulation research, remains applicable to all programs.

References

Bejar, I.I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283-296.

Bejar, I.I. (1988). An approach to assessing unidimensionality revisited. *Applied Psychological Measurement, 12,* 377-379.

Bloom B. S. (1956). *Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain.* New York: David McKay Co. Inc.

Bock, R., Gibbons, R., Schilling, S., Muraki, E. Wilson, & Wood, R., (2003) TESTFACT 4 (Computer software). Lincolnwood, IL: Scientific Software International.

Castro, F., Suarez, J., & Chirinos, R. (2010). *Competence´s initial estimation in computer adaptive testing.* Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Choi, S.W. (2009). Firestar: Computerized adaptive testing (CAT) simulation program for polytomous IRT Models (Computer software). *Applied Psychological Measurement, 33,* 644–645.

De Ayala, R.J. (2009). *The theory and practice of item response theory.* New York: Guilford.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23,* 249-261.

Eggen, T. J. H. M, & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60,* 713-734.

Embretson, S.E., & Reise, S.P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Flaugher, R. (2000). Item Pools. In Wainer, H. (Ed.) *Computerized adaptive testing: A Primer.* Mahwah, NJ: Erlbaum.

Frick, T. (*1992*). Computerized Adaptive Mastery Tests as Expert Systems. *Journal of Educational Computing Research, 8(2),* 187-213.

Georgiadou, E., Triantafillou, E., Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Retrieved August 23, 2010 from http://www.jtla.org .

Gibbons, R.D., Weiss, D.J., Kupfer, D.J., Frank, E., Fagiolini, A., Grochocinski, V.J., Bhaumik, D.K., Stover, A., Bock, R.D., Immekus, J.C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*, 361-368.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Kolen, M.J., & Brennan, R.L.. (2004) *Test equating, scaling, and linking. Methods and practices,* 2nd ed. New York: Springer.

Lee, J, & Weiss, D. J. (2010). *Selection of common items in full metric calibration for the development of CAT item banks.* Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.

Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.

Nydick, S. W., & Weiss, D. J. (2009). A hybrid simulation procedure for the development of CATs. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved August 23, 2010 from www.psych.umn.edu/psylabs/CATCentral.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2006). *Practical considerations in computer-based testing.* New York: Springer.

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 August 23, 2010 from http://www.psych.umn.edu/psylabs/CATCentral .

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Reckase, M. D. (2003). *Item pool design for computerized adaptive tests.* Paper presented at annual meeting of the National Council on Measurement in Education, Chicago, IL.

Rudner, L. M. (2002). *An examination of decision-theory adaptive testing procedures.* Paper presented at the annual meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA.

Rudner, L.M. and Guo, F.M. (in press) Computer adaptive testing for small scale programs and instructional systems. *Journal of Applied Testing Technology.*

Sands, W.A., Waters, B.K. and McBride, J.R. (Eds.) (1997). *Computerized adaptive testing. From inquiry to operation.* Washington: American Psychological Association.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation*, 12(1). Available online: http://pareonline.net/getvn.asp?v=12&n=1 .

Thompson, N.A. (2009a). Item selection in computerized classification testing. *Educational and Psychological Measurement, 69*, 778-793.

van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing*, (Statistics for Social and Behavioral Sciences Series). New York: Springer.

van der Linden, W.J. (2010). *How to make adaptive testing more efficient?* Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Veldkamp, B.P., & van der Linden, W.J. (2010). Designing item pools for adaptive testing. In van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing*, (Statistics for Social and Behavioral Sciences Series). New York: Springer.

Verschoor, A. (2010). *Optimal calibration designs for computerized adaptive testing.* Paper presented at the first annual conference of the International Association for Computerized Adaptive Testing, Arnhem, The Netherlands.

Vispoel, W.P., Rocklin, T.R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized adaptive, and self-adapted testing. *Applied Measurement in Education, 7*, 53-59.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice, 12*, 15-20.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A Primer.* Mahwah, NJ: Erlbaum.

Waller, N. (1997). *MicroFACT* (Computer software). Saint Paul, MN: Assessment Systems Corporation.

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

Weiss, D. J. & Guyer, R. (2010). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Assessment Systems Corporation.

Weissman, A. (2004). *Mutual information item selection in multiple-category classification CAT.* Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Diego, CA.

Wise, S. G., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologia, 21*(1), 135-155. http://redalyc.uaemex.mx/pdf/169/16921108.pdf

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model.* Saint Paul, MN: Assessment Systems Corporation.

Yoes, M. (1997). *PARDSIM* (Computer software). Saint Paul, MN: Assessment Systems Corporation.

Citation:

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.

Corresponding Author:

Nathan A. Thompson, Vice President
Assessment Systems Corporation
2233 University Ave., Suite 200
St. Paul, MN. 55114

Phone: +1 (651) 647-9220
Email: nthompson [at] assess.com