Series Editor's Note

David Weiss, the father of computerized adaptive testing (CAT), with his coauthor, Alper Şahin, bring everything I had hoped for in this comprehensive and readily accessible guide to all things CAT. Weiss is a font of wisdom in the world of CAT. His voice is *the* voice, steeped in practical advice born of the tens of thousands of hours applying the principles of CAT to a wide array of research problems. The skill he has acquired by developing, innovating, and applying CAT is clear in the eloquent exposition of this unequivocally great work that he and Şahin have created. It's a true how-to guide to the craft of CAT.

This book is an indispensable tour de force, unlike any other book on CAT. The authors cover the ins and outs of item response theory, item calibration, proficiency estimation, exposure control, content balancing, item bank development, and the like, in a manner that is nontechnical and thorough. For professionals in measurement and assessment, this book will be your go-to source for implementing an accurate instrument that will provide a unique test for each person. CAT provides this personalized assessment in real time by utilizing the latest innovations in artificial intelligence and machine learning. This book provides everything you need to know to develop and implement an accurate and efficient assessment that won't tax test takers. Instead of testing sites and proctors, test takers can use any internet-enabled device at a convenient time and place to provide the information needed for making an accurate assessment.

In an era of high-stakes testing, precision is paramount. CAT is the only testing platform that offers the potential for high-stakes precision in educational assessment and professional certification. Using a variety of examples where precision is critical, this resource is for anyone who needs either to examine person-level dif-

ferences, including individual change, or to group and classify persons as needed. The insights, born of experience and knowledge, for planning and creating an item bank that is tailored to your specific application are indispensable. The various software platforms and background essentials are detailed, allowing anyone to build and implement a CAT from start to finish.

As always, enjoy!

Todd D. Little
At my "Wit's End" in Lakeside, Montana

1

Why CAT?

WHAT IS CAT?

Computerized adaptive testing (CAT) is the redesign of tests of ability, achievement, aptitude, proficiency, personality, preferences, attitudes, and other psychological, educational, or human resources variables for administration by a computer. But it is not simply a conventional test—in which the items and their sequence are fixed in advance—that is administered on a computer. Contemporary CAT uses principles of artificial intelligence and machine learning to design a test for each examinee while test administration is in progress, scoring each answer as it is provided. Thus, as the examinee answers the test items, the artificial intelligence algorithm does what a trained psychologist would do—it selects test items for each examinee based on their answers to questions the examinee has already answered and immediately scores their answers. In the process, the algorithm continuously "learns" what the examinee's trait level is and adapts item selection to that examinee. Consequently, each examinee has the potential to receive a unique test. As a result, CATs have psychometric benefits that impact examinees, other benefits to examinees, and benefits to the testing organization. At the same time, CATs raise some new challenges to both examinees and testing organizations. This chapter introduces these benefits and challenges and some feasibility issues. In doing so, it presents an overview of much of the material that follows in many of the chapters.

BENEFITS OF CAT

Psychometric Benefits

Since the early 1900s, psychological measuring instruments of all kinds have been developed and implemented using some version of "true + error" score theory (Gulliksen, 1987). This classical test theory (CTT) focuses on internal consistency

reliability as the primary objective for test construction. In CTT, reliability (r_{xx}) is defined as the ratio of true score variance to error variance and is frequently operationalized as coefficient alpha (α; Cronbach, 1951) or Kuder-Richardson formula 20 (Kuder & Richardson, 1937) for dichotomously scored (correct/incorrect, keyed/nonkeyed, true/false) items,

$$r_{xx} = \alpha = \mathrm{KR}_{20} = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} p_i q_i}{V_x}\right) \quad (1\text{-}1)$$

In Equation 1-1, n is the number of items in the test, p_i is the proportion of examinees who correctly answered the item, $q_i = 1 - p_i$ (i.e., the proportion who answered incorrectly), and V_x is the variance of the total (number-correct or sum) scores. According to CTT, algebraic manipulation of this equation results in two criteria for maximizing internal consistency reliability: (1) select items with proportions correct around $p_i = 0.50$ and (2) select items with high correlations with total scores (i.e., item discriminations). Thus, procedures of item analysis based on CTT will eliminate items for which the proportion correct deviates from approximately $p_i = 0.50$ and will also eliminate items with low discrimination. The result is a set of highly discriminating items with difficulties around $p_i = 0.50$, and a test with high reliability. High reliability translates into high precision and low standard error of measurement.

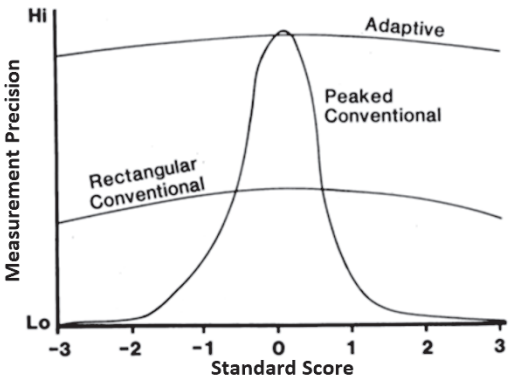
However, when viewed from the perspective of modern test theory—namely, item response theory (IRT), as described in Chapter 3—tests constructed using the principles of CTT have suboptimal properties for measuring individuals. CTT reliability is a group statistic. A reliability coefficient can be computed from the administration of a fixed set of items to a specified group of individuals, and the result is a single value. This implicitly assumes that the reliability—or precision—of test scores is constant for all examinees. But IRT permits the examination of measurement precision conditional on score level. CTT reliability also is group dependent and should be computed for every administration of a test to every group of examinees for which it is used. This dependency is obvious in Equation 1-1: A larger score variance (V_x) will increase r_{xx} and a larger sum of the item variances ($p_i q_i$) relative to score variance will decrease r_{xx}.

A test constructed according to CTT principles can be termed a *peaked test* because the distribution of item difficulties is concentrated around $p = 0.50$—a frequency distribution of item difficulties would show most items with difficulties around $p = 0.50$, with smaller numbers of items with difficulties that deviate from 0.50 in either direction. There will likely be virtually no items with, say, $p = 0.20$ or lower or $p = 0.80$ or higher. When viewed from the perspective of IRT, however, the conditional precision of a peaked test will resemble the function shown in the

center of Figure 1-1. The test will measure very precisely for individuals of average trait level, but precision will fall off drastically as person trait level moves away from average. Consequently, for examinees with trait levels more extreme than one standard deviation below (–1) or above (+1) the mean, scores will have virtually no precision and represent essentially random error. This occurs because for those examinees there are no items in the test that are appropriate for their trait level—the items are too difficult for examinees with low trait levels and too easy for those with high trait levels. For examinees with scores around average, their proportion correct on the test will be near the optimal of 0.50, but for low trait level examinees, the proportion correct will converge on 0.0, while for high trait examinees, the proportion correct will converge on 1.0. In both of the latter two cases, scores will be unable to reflect individual differences as they exist among the examinees, and measurement precision will converge on 0.0.

Although not formally supported by CTT, conventional fixed-form tests can be constructed as *rectangular tests*, based on their distribution of item difficulties across the full range from $p = 0.0$ to $p = 1.0$, with a reasonably equal number of items across the difficulty range, but still select items with high discriminations. Because the test has a fixed total number of items, this kind of test will have a few items for very low trait examinees, a few for very high trait examinees, and a relatively equal number of items for all trait levels. However, overall reliability will be lower for this test design than for a peaked test of similar length, resulting in decreased precision overall, but the precision will be relatively equally distributed across the score range, as shown in Figure 1-1. Thus, in contrast to the peaked test, the rectangular test will measure individuals equally well across the trait, but precision for examinees in the center will be lower than that of a peaked test, yet the remaining examinees will be more precisely measured.

FIGURE 1-1. Measurement precision of peaked, rectangular, and adaptive tests as a function of trait level.



Equal Measurement Precision

Figure 1-1 also shows the precision function for a well-designed and well-implemented CAT. As shown in Chapter 2, the Binet adaptive test selects, from a precalibrated item bank, a set of items for each examinee that results in approximately 50% correct answers. From a psychometric perspective, this is the set of items that provides the most information about an examinee—and this information results in maximum precision. Because the adaptive test selects a tailored set of items for each examinee, adapting to their trait level as the test is administered, an adaptive test can provide high measurement precision for all examinees, regardless of trait level. For examinees of average trait level, measurement precision can be the same as that of a peaked conventional test, but that same level of precision will be possible for all examinees. This results as the adaptive procedure dynamically creates a "peaked test" that is unique to each examinee.

Improved Measurement Efficiency

CAT makes the testing process efficient in this way by using the examinee's testing time and the items more efficiently. Each examinee is presented questions for which they have a predicted probability of around 0.50 of responding correctly during the test because the items are chosen from a precalibrated bank of items with a difficulty level closest to the estimated ability/trait level of the examinee during testing and, within that subset of items, items that are most discriminating at that ability level (Weiss, 1983). Consequently, CAT psychometrically tailors the test to the ability level of the examinee, resulting in efficient use of the examinee's testing time and increases in measurement precision (Wainer, 2000; Weiss, 1973).

Many studies over more than a half-century of CAT have shown that CAT can decrease test length by at least 50% without sacrificing measurement precision (e.g., Kimura, 2017; Olsen et al., 1986; Schnipke & Reese, 1997). Gibbons et al. (2008) applied IRT-based CAT to a 616-item psychiatric instrument that measured mood and anxiety and observed an average 95% reduction in the number of items in both post-hoc simulations (see Chapter 9) and live testing. In live testing, an average of 30 items resulted in a correlation of 0.93 with scores based on all the items, and average test administration time was reduced from 115 minutes to 22 minutes.

Person-Specific Error of Measurement

All measurements involve some degree of error. A major focus of psychometrics is to measure each person with the lowest amount of error possible. As described above, in fixed-form tests developed by CTT, a single value of error—the standard error of measurement or SEM—is calculated from the reliability coefficient by

$$\text{SEM} = \text{SD}_x\sqrt{1 - r_{xx}} \quad (1\text{-}2)$$

where $\text{SD}_x = \sqrt{V_x}$. Because both SD_x and r_{xx} are single values that result from the administration of a single test to a specific group of examinees, SEM will also be a single value for that group, implying that all examinees are measured with the same degree of (im)precision. However, in reality, the SEM might dramatically differ among examinees, and it is not possible to identify those differences with CTT. Powered by IRT, CAT brings this important psychometric benefit to the testing process. Using CAT, the trait/ability level of each examinee can be estimated with the same or similar controlled amount of error as operationalized in an SEM explicitly conditional on trait level. Moreover, an SEM can be calculated for each examinee individually (de Ayala, 2022), based only on the examinee's pattern of responses to the test items they answered and the psychometric characteristics of those items (see Chapter 3). In this way, the precision of the trait/ability estimates is based on the same metric for all examinees, can be compared and controlled using CAT, and can vary from examinee to examinee (see Chapters 4 and 5).

Benefits for Examinees

Guessing has been a problem since the beginning of standardized testing, as most standardized tests use multiple-choice item formats. CAT can help reduce the effects of guessing in three ways. First, thanks to IRT's three-parameter logistic model (see Chapter 3), guessing is one of the item parameters calculated for each item, in addition to the two other item parameters: difficulty and discrimination (but not the difficulty and discrimination of CTT). The guessing parameter is included in the model equation and is considered during the ability/trait estimation process. Second, in traditional fixed-form tests, examinees (particularly those of low ability) tend to guess the correct answer when they are presented items that are above their ability level. However, as mentioned earlier, in CAT test items are aligned to each examinee's ability/trait level. This can reduce or eliminate guessing behavior during CAT sessions for low-ability examinees. Finally, because the test is being delivered on a computer, item types that can be used in CATs are not limited to multiple-choice items. Examinees can be asked to type or speak their answers or they can be asked to order the choices, among other creative modes of answering test items that can be immediately scored by computers.

A well-designed CAT focusing on the measurement with individual differences (Chapter 3) draws items from an item bank that is designed to measure examinees to the same level of precision (low SEM) regardless of their measured trait level. The result of this objective is a test in which each examinee, regardless of their ability/trait level, receives a set of items for which they answer approximately 50% of the items correctly. This contrasts with fixed-form conventional tests in which many examin-

ees typically are required to answer items that might be well below or above their actual ability level. This sometimes works as a demotivator and can cause frustration during the test for examinees with low ability levels, or a source of boredom for those with high ability levels. CAT methodology eliminates this problem by presenting each examinee with items tailored or adapted to their estimated ability level as it is continuously estimated during the test session. In this way, even the student with the lowest ability level will be able to answer items suited to their ability and they can feel less anxious responding to items that are not overly frustrating for them. Similarly, examinees with high ability levels will be administered items that are still challenging for them, thus potentially motivating them to pay closer attention to the test. Thus, a CAT that is well developed and properly delivered can ensure that the items are used efficiently, and examinees are given items that are of a level of difficulty appropriate for each person (Green, 1983; Sands et al., 1997), potentially equalizing the psychological environment of the test-taking process for all examinees.

Some high-stakes fixed-form tests are still administered to large groups of examinees at various locations. Answer sheets are then collected and shipped to a scoring service. It then can take as many as 15 to 30 days to report test results, including examinee scores. This is due to such factors as the physical transfer of the test documents, transfer of the answer sheets for scoring, scanning of the test answer sheets, analysis and preparation of results, and shipment of results to the testing organization. However, CATs provide the possibility of providing the examinee with instant, at least preliminary, results immediately after the test is completed (Wainer et al., 1990).

Moreover, some high-stakes testing programs allow the examinee to accept the test result or to cancel it immediately after seeing their score following the test session. An examinee who cancels a test score immediately after a test session can decide to take the test again as soon as possible, depending on the test publisher's policy for allowing retakes. This is an invaluable benefit supplied to the examinee by CAT. Typically, if it was a test delivered in paper-and-pencil format, the examinee would need to wait until a new test window is opened for another session of the test; however, CAT eliminates this limitation and gives the examinee the freedom to take the test at a time that they wish (Patsula, 1999). Knowing that they can take the test at a time and place convenient for them on their first attempt and that there will be an opportunity to take the test multiple times will also ease the tension on the examinees and possibly decrease examination anxiety (Glas & Van der Linden, 2001).

As indicated, CAT requires fewer items compared with traditional fixed-form tests. This makes CAT much more efficient in terms of test duration. This feature of CAT directly affects examinees because receiving fewer items in a test means that they need to spend less time responding to those items. This is undoubtedly advantageous to the examinee, who must spend long periods of time on fixed-form tests, and should decrease the effect of fatigue on examinee scores. Although there might still be CAT sessions that last an hour or two for measuring multiple variables (although

multidimensional CAT can further increase efficiency; see Chapter 12), there are CATs that are as short as 30 minutes and others measuring medical outcome variables that obtain precise scores measuring three patient-reported outcomes variables in an average of 6 minutes (Wang et al., 2022). As mentioned previously, Gibbons et al. (2008) applied CAT to a psychiatric inventory and reduced average administration time by over 90 minutes. Clearly, CAT can dramatically decrease the time that examinees spend taking tests.

Organizational Benefits

The item types that are used in traditional paper-and-pencil tests are obviously those that can be administered on paper and responded to on paper. This results in many limitations that affect the item writing process. For example, in paper-and-pencil tests, examinees cannot watch a video and then respond. CAT allows the implementation of many kinds of technology-enhanced items such that item writers and test publishers can benefit from the use of technology as much as possible with the help of computers to move testing to a more realistic type of test item, thus allowing the measurement of many kinds of traits and abilities that are not otherwise measurable (see Chapter 15).

It is not always possible to guess or estimate the psychometric quality of an item before it is administered to real examinees in a real test session. If the psychometric quality of an item is not determined, it means that the psychometric quality of the tests in which the item might be used cannot be determined. For this reason, there is a need for a testing organization that publishes standardized tests over a period of time to pilot some newly written items before they are operationally used in real tests. This has sometimes been a difficult process if fixed-form tests are used in such testing programs. However, CAT brings another psychometric benefit to the industry for piloting of some newly written items by "seeding" these items among the actual operational items during CAT test sessions (see Chapter 8). Sometimes the examinees are informed before the test that they might have some experimental items interspersed in their CAT but that they will not be counted in the scoring of the test. However, examinees are not informed about which items are operational items and which are experimental items.

In traditional paper-and-pencil tests, a company incurs certain costs on a continuing basis, including careful proofreading of the final printed tests, printing the booklets, storage of the printed booklets before and after the test, purchase of answer sheets, costs of shipping both booklets and answer sheets, the need to store the booklets and answer sheets in secure rooms, the costs of scanning, and the costs of reporting test results. Such costs are not borne by a company administering tests in CAT format (Rudner, 1998).

In high-stakes applications of educational testing, such as college and university admissions, cheating scandals are one of the main problems that testing organiza-

tions have faced in recent decades. If paper-and-pencil tests are being used, there is an increased probability that a test could be compromised. CAT uses algorithms that select the items for each examinee on an "as-needed" basis based on the continuously estimated trait level of each examinee, solely depending on the responses that each examinee provides during test delivery. This means that there is almost no possibility that a test delivered in CAT format can be stolen in advance. Even if the item bank being used for the CAT was stolen, there would be virtually no chance of knowing which items would be used in the test administered to any specific examinee. Thus, in CAT, because each examinee receives a different set of questions/items, and items are selected from a relatively large bank of items that can be securely encrypted in the memory of a central computer operating behind electronic firewalls, there is no need to physically secure printed copies of the items, and obtaining some of the questions in the bank before the test is delivered to a given examinee will likely have little or no impact on test scores (Green, 1983; Patsula, 1999; Thompson, 2011a). This enhanced level of security is one of the most important benefits of CAT for testing organizations.

Concerning the security-related benefits of CAT, one of the reasons for the early attempts to switch paper-and-pencil tests to CAT in some well-known testing programs in the United States (e.g., the CAT-ASVAB, used for testing military recruits; see Chapter 11 for more details) was the need to test candidates without physically transporting them to test centers, accommodating them, and feeding them (McBride, 1982). Before ASVAB was delivered in CAT format, candidates had to take long tests in paper-and-pencil format. However, today CAT allows for online testing with remote proctoring through the Internet. There are many options for remote proctoring. Almost all options lock the computer system and do not let the examinees browse on the Internet or execute any software other than the testing software, while digitally recording the examinee during the test session. Another option, typically used in testing for hiring potential employees, allows the examinees to take the test online with no or partial restrictions and then asks them to later take a short confirmation test covering the same material under supervision. Other options include live remote monitoring of test administration through the examinee's webcam and recorded monitoring with a post-testing review of the recording (further discussed in Chapter 6).

Another significant benefit that CAT provides to organizations is the opportunity to reach examinees online worldwide while maintaining the security of their tests. This capability might also increase examinee satisfaction as they will not need to travel long distances to take their tests. This capability would also be especially beneficial for universities that accept applications for their programs worldwide as well as multinational organizations.

An organization that uses a paper-and-pencil testing program for large numbers of individuals needs staff for proctoring/invigilating tests during test sessions. However, implementing CAT methodology in the testing program will reduce the

demands on the invigilation/proctoring staff as CAT allows for on-demand testing and, as indicated, several types of electronic proctoring. Examinees can take some tests whenever and wherever is convenient for them, and they do not necessarily need to complete the test on a specific date. This decreases the test administration staffing needs of the organization and also possible errors caused by human staff members during delivery, such as incorrect timing of testing time.

Finally, switching a currently active testing program to the most cutting-edge test technique in the industry, CAT, would likely increase an organization's value in the eyes of its clients, especially if both the psychometric benefits and the other benefits to examinees are carefully factored in. The global testing market is huge. There are many competitors in the field of testing, and in order to achieve a prominent place in this competition, it might be helpful to switch testing programs to CAT. Not doing so could negatively affect an organization. By switching tests to CAT format, the first benefit that will be realized by an organization is to advertise that it is using the latest and most accurate techniques in testing, and this should increase the visibility of the testing organization.

CHALLENGES OF CAT

Although CAT brings many benefits and innovations to testing organizations, examinees, and the psychometric quality of test scores, there are also some challenges that need to be dealt with by the organizations, examinees, and psychometricians using CAT.

Organizational Challenges

The first challenge faced by organizations planning to implement CAT is the funding needed to invest in a computer system that will enable item banking, test assembly and delivery, item scoring, estimating and using IRT item parameters, and software to manage the bank and the resulting examinee data. CAT administration needs specifically designed software that will house the item bank and an item writing workbench to manage the flow of item development. This software must allow item writers and item reviewers to create and access test items, enable the assembly of tests, and specify the options under which a particular CAT will function. Moreover, this software optionally and ideally should be able to deliver items in the bank directly in the form of CAT. The software should also provide for reporting of examinee and group test results, item statistics, and various forms of data analysis.

Coding and checking for the accuracy of the performance of such a system will be costly and time-consuming. In many cases, it will likely be more cost-effective to use a ready-made commercial, professionally developed CAT software that will allow you to create, edit, and bank your items, design your CATs, and deliver and

report them, as well as analyze your items and test results (e.g., FastTest, which is briefly described in Chapter 10). Another option might be to use open-source free platforms (e.g., Concerto, which is also briefly described in Chapter 10) for this purpose. However, open-source software might be less reliable for important CATs and will likely require a degree of sophistication to install and connect to the web.

Purchasing continuing access to a professional CAT platform, or building one for a specific CAT application and setting up the computer systems that will be responsible for the delivery of the test, will not be enough to deliver your CAT tests. This issue is another challenge an organization will face while transforming their tests to CAT: It is the need to have in-house expertise on IRT and CAT (Thompson, 2011a). Contemporary CAT is based on IRT, so an additional cost is the hiring of a professional or professionals who is/are CAT and IRT expert(s). Such expert support might be continuously required, as in some cases the item bank will need to be updated regularly. These costs should also be included in the CAT budget as additional costs for setting up an optimally functioning CAT delivery system.

Once a CAT-friendly system and expert support have been put in place, it is then time to think about the content of the test. This might not be a challenge for organizations already administering their tests on computers because they already have the online versions of the items. The main issue they need to face is to find a way to transfer their existing item database to the CAT item bank. However, if the testing program is migrating its tests from paper-and-pencil format to CAT, to administer the tests using formats amenable to online item delivery and obtain faster item analysis and score reporting, they then have the opportunity to convert their items to new item types. Under these circumstances, of course, it will be necessary to design and empirically try out the new item types with the test's target population. The next phase would then be assembling a CAT item bank with items calibrated by an appropriate IRT model (see Chapters 3, 7, and 8).

Another organizational challenge that might be faced when switching to CAT is the need to provide a secure computer system that has the item bank. The item bank is the core of CAT. If the computer system that houses the CAT item bank is hacked or if an instance of a security breach occurs, this type of situation could sabotage all the efforts that were invested in item writing, item analysis, and item bank development. Moreover, the exam scores, examinee data, and CAT session details will also be stored on this server; these are categorized as sensitive data, and their loss or theft can cause problems for the testing organization. Consequently, the level of security of the computer or the server that stores and delivers the CAT should be taken seriously.

Examinees who are used to taking tests with their friends might find their CAT experience a bit unusual. A well-designed and well-implemented CAT will use a variable termination rule, likely resulting in different test lengths for different examinees (but potentially equal measurement precision or accuracy). Thus, one examinee might answer 30 questions, while a friend or colleague taking the same test at

the same time might answer only 22 questions. This might cause the examinees to conclude that the test is unfair. CAT design, test delivery options, and quality should be carefully explained to the examinees; this issue is briefly addressed in Chapter 13.

A secure CAT delivery system, expertise in IRT and CAT, and the ability to educate the examinee about CAT will not be enough to control all the challenges that an organization can experience. There is also a need for careful documentation of the CATs administered. This documentation will be needed for public review, and release of white papers and other technical documentation for the CAT might be necessary. Otherwise, an organization could experience legal issues in some countries, depending on the laws in that particular country. Therefore, both the public's and the examinee's perception of a CAT program should be also taken as seriously as the other organizational challenges described above.

Examinee Challenges

Tests administered in CAT format not only entail some challenges for testing organizations, but they also entail some challenges for the examinees. First, in tests administered in CAT format, examinees are constrained to answer one question at a time. Moreover, they have no chance to review the items they answered previously (Bakker, 2014). Consequently, they have no opportunity to change their answers because each new item presented to an examinee is determined by the completed pattern of correct and incorrect responses provided by the examinee for all previously administered items. The lack of the ability to change the responses to the previous items is one of the most problematic perceived disadvantages of CAT for examinees (Lunz et al., 1992; this issue is further discussed in Chapter 13).

Another challenge that CAT examinees face is that they typically have to answer all questions in the test; otherwise, they might either be penalized if unanswered questions are considered incorrect, or the question is simply not scored. In most cases, an unanswered item for an ability or achievement test (if allowed by the CAT designer) will likely be scored as incorrect—on the assumption that it was not answered because the examinee did not know the correct answer—since the CAT algorithm calculates the scores according to the correct or incorrect responses given to each item. If, however, the CAT designer allows items to be skipped, the algorithm will simply select the next best item at the current ability level, thus ignoring the skipped item but increasing the test length.

Another potential disadvantage of CAT for examinees is the inability to see all the items at once and choose which item to begin the test. In a traditional paper-and-pencil test, the examinee can see all items at once and is free to start the test with whichever items are preferred. This is simply not possible in CAT due to its adaptive nature. The examinee needs to answer the items on the screen as they are presented and has no choice over the order in which items are answered.

Psychometric Challenges

The psychometric challenges that a CAT entails begin with creating an up-to-date item bank with psychometrically sound items. The need to maintain the item bank with new items, and retire some older items, while maintaining all past and future score estimates on the same scale will also be another challenge in some testing programs. This will require the item writing and item analysis process to be continuous. New items are evaluated during a pilot phase and must be linked to the original score scale before they are put into operation. At this stage, depending on the IRT model used, appropriate sample sizes and linking/equating procedures need to be used to complete the process of adding items to the bank; these issues are discussed in Chapters 3 and 8, respectively, while Chapter 7 discusses some nonpsychometric aspects of item bank development.

Because the item bank is the core of CAT, a psychometric challenge that awaits organizations and researchers is that of creating, maintaining, and using an item bank that has items that provide information covering a wide range of ability/trait levels. A well-designed CAT is frequently intended to end when a prespecified level of precision, as operationalized by the SEM, is reached. If an item bank has few items available at some trait levels—usually very high and very low—the CAT algorithm will select items that are less than optimal for those examinees, resulting in longer tests and tests that in some cases might not be able to be terminated to the prespecified level of SEM. This issue of suboptimal item banks is addressed in Chapters 4 and 5.

A psychometric challenge that arises in some CAT applications—most notably achievement tests—is the differences in test content for some examinees due to a relationship between test content and item difficulties in different CAT administrations. For example, one examinee's CAT might present 10 algebra questions and no geometry questions, whereas another examinee might get 10 geometry questions and no algebra questions. This is a challenge that unconstrained CAT users face under some circumstances. In order to overcome this difficulty, some solutions for content balancing have been developed (see Chapter 13 for information on content balancing). Although content balancing can solve this problem, it becomes the source of another problem because it requires having more items in the CAT bank for different types of content. Development, piloting, and calibration of these items is an additional psychometric challenge for CAT developers.

Depending on the number of items and the number of examinees taking a CAT, some items might be used more frequently than other items in the item bank. Such items are called *overexposed items*, which can be a problem in high-stakes tests that are administered on demand. In this test setting, it might be appropriate to retire overexposed items after a period of time, replacing them with new items of similar psychometric quality. Item overexposure is another challenge that some CAT item bank developers and psychometricians face. Chapter 13 also discusses issues of item

exposure and how to control it in circumstances in which it might affect test integrity.

A psychometric problem that occurs in some CAT applications for some examinees occurs when an examinee responds to a set of questions with all incorrect or all correct responses. IRT scores are frequently estimated using a method called maximum likelihood estimation (MLE). As described in Chapter 3, MLE cannot calculate a score for examinees who answer all items in a test correctly or incorrectly. Although this is a rare event in a well-designed CAT—it typically occurs for examinees with extremely high or low ability/trait levels—Chapter 13 describes and recommends methods that can be used to avoid this problem.

WHEN IS CAT FEASIBLE AND WHEN IS IT NOT?

The feasibility study necessary before implementing a CAT is introduced in detail in Chapter 6. In Chapter 6, some aspects of CAT that constrain its practicality, cost, and the time/effort needed to implement a viable CAT testing program are considered.

The most common item type used in CAT is multiple-choice items. In addition to multiple-choice items, ordered choice, constructed response, and forced-choice item types can also be used. CAT is currently unable, with a few exceptions, to score constructed response items that require the examinee to respond to a question with more than a phrase. The reason for this is so that the CAT system is able to objectively determine in real time whether the response of the examinee is correct or incorrect—or assign a score based on an evaluation of the quality of the response—in order to estimate the ability/trait level and continue the testing process. CAT is an early implementation of artificial intelligence (AI) in some aspects of its operation, and some testing organizations (e.g., ETS, Pearson) have already developed AI engines to score essays. But AI scoring of essays is currently not in use in CAT, so if the ability to be tested requires subjective scoring of written paragraphs or long spoken responses, it is not currently feasible to use CAT to test these skills. For now, the items in the CAT item bank should be those that can be scored immediately and objectively by the computer as correct or incorrect (or use rating scales to measure noncognitive traits), although advances in AI should, in the near future, allow items that require scoring by processing natural language—whether it be written or spoken—to be used in CAT applications.

As mentioned earlier, items in some CAT item banks should be maintained continuously (McCloy & Gibby, 2011). This process means that new items need to be added to the item bank and some old items need to be retired or recalibrated. As indicated above, it is possible to insert new items into an operational CAT interspersed among the operational CAT items—a process called "seeding" (see Chapter 8). However, this requires a certain number of examinees to respond to these pilot items in order to obtain the necessary number of respondents to accurately

estimate the item parameters of the newly written items. If the number of examinees in the program is very limited (e.g., less than 300), it might not be feasible to implement CAT if the testing program requires continuous replenishment of the item bank because (1) there are insufficient numbers of examinees to estimate the item parameters needed to implement the CAT (see Chapter 3) and (2) new items to be added to a bank require at least that many examinees to estimate their item parameters.

CAT provides its users with high levels of score precision and efficient use of examinee time in return for a substantial initial effort and investment. The result, however, is a measurement system that can be scientifically and legally defensible. CAT is especially important in high-stakes test applications where important decisions are made about the examinees. In these applications, the cost and effort to implement CAT are worthwhile. However, if the plan is to use CAT as a classroom test for 20 students, it would not be feasible to set up a CAT system, unless the CAT was developed by a central authority in the school system and made available to classroom teachers. This is an approach that has been taken in many large-scale school systems, and in some states at the state level, both for purposes of the diagnostic evaluation of students to guide instruction and for evaluating teaching approaches and practices. Chapter 11 provides examples of diagnostic CAT that are successfully in use around the world, and Chapter 12 includes a brief technical discussion of how a special type of CAT is being implemented for evaluating student mastery of specific academic skills.

In a testing situation, all examinees should be provided equal opportunities and necessary precautions should be taken to provide all examinees with the same or similar test environment. For group CAT administrations, this can mean that monitors, computer hardware, and the Internet connection must be high quality for all examinees so that they do not negatively impact the testing process. Testing room conditions should also be reviewed. For example, the monitors used should have an anti-glare coating and should be optimized for the best view of graphics and text (Wainer et al., 1990) and include the necessary capabilities to adapt appropriately to the testing of students with disabilities.

Because power failures can occur, either for an entire testing room or individual testing computers, a backup software system and backup power supply should also be in place for the testing room. Also essential is a CAT software delivery system that allows each student to recover and continue from the last item administered in the event of failure due to power, hardware, software, Internet connection, or other unanticipated events (e.g., sudden illness of an examinee). Thus, it is most important that test conditions are reviewed carefully to provide the examinees with a positive testing experience, free from extraneous factors that can negatively affect examinee test scores. If such precautions cannot be taken, CAT might not be feasible in that test environment.

Despite all the challenges, the advantages of CAT have prevailed in many testing programs around the world. From its beginnings in educational and military testing in the early 1970s, CATs can be found in virtually every major country and are being used in many applications that go well beyond these early applications. Chapter 11 provides some examples of operational CATs in a variety of applications, with the most recent in healthcare permitting rapid and precise measurements of patient-reported symptoms, with many new CATs being developed to measure a wide variety of personality and other psychological variables.

Guilford Publications
370 Seventh Avenue
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com